UMA INVESTIGAÇÃO A RESPEITO DE SISTEMAS DE ATENÇÃO VISUAL COMO EXTRATORES DE CARACTERÍSTICAS LOCAIS

ALANA DE SANTANA CORREIA

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, DEL, da Universidade Federal de Sergipe, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Eduardo Oliveira Freire

Sergipe
Fevereiro de 2019

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA-PROEE

TERMO DE APROVAÇÃO

"Uma investigação a respeito de sistemas de atenção visual como extratores de características locais"

Discente:

Alana de Santana Correia

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Sergipe, como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica.

Aprovada pela banca examinadora composta por:

Prof. Dr. Jugurta Rosa Montalvão Filho (PROEE/UFS)
Presidente

Profa. Dra. Beatriz Trinchão Andrade (DCOMP/UFS)
Examinadora Externa

Prof. Dr. Jânio Coutinho Canuto (PROEE/UFS)
Examinador Interno

Alana de Santana Correia
Candidata

Cidade Universitária "*Prof. José Aloísio de Campos*", 1 de fevereiro de 2019.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Correia, Alana de Santana
C824i Uma investigação a respeito de sistemas de atenção visual como extratores de características locais / Alana de Santana Correia ; orientador Eduardo Oliveira Freire . - São Cristóvão, 2019.
130 f. : il.

Dissertação (mestrado em Engenharia Elétrica) – Universidade Federal de Sergipe, 2019.

1. Engenharia elétrica. 2. Sistemas biológicos. 3. Percepção visual. 4. Atenção. 5. Inteligência artificial. I. Freire, Eduardo Oliveira orient. II. Título.

CDU 621.3:616

Resumo da Dissertação apresentada ao PROEE/UFS como parte dos requisitos necessários para a obtenção do grau de Mestre (M.Sc.)

UMA INVESTIGAÇÃO A RESPEITO DE SISTEMAS DE ATENÇÃO VISUAL COMO EXTRATORES DE CARACTERÍSTICAS LOCAIS

ALANA DE SANTANA CORREIA

Fevereiro/2019

Orientador: Eduardo Oliveira Freire

Programa: Engenharia Elétrica

A atenção é um sistema biológico responsável pelo gerenciamento de recursos dos seres vivos. É por conta do mecanismo de atenção que eles conseguem realizar tarefas, coordenar as atividades motoras, e até mesmo perceber o ambiente. No que diz respeito à percepção, existe a atenção visual, que filtra a grande quantidade de informações visuais captada do ambiente. Somente os elementos perceptualmente mais importantes são armazenados na memória e utilizados pelo sistema cognitivo no reconhecimento e recuperação de informação. Inspirada pelas descobertas da atenção visual biológica, se desenvolveu a atenção visual computacional, uma área destinada a construir sistemas atencionais em máquinas. Além dessa área, também se desenvolveu a área de extração de características locais, que tem como objetivo encontrar e descrever pontos que sejam resistentes às transformações básicas em uma imagem, como escala e rotação, por exemplo. Apesar de serem idéias provenientes de inspirações biológicas obtidas a partir de observações do sistema de atenção visual humano, essas áreas produzem ferramentas computacionais bem distintas. Já existem alguns esforços por parte dos pesquisadores em realizar uma junção das ferramentas de atenção com as ferramentas de extração, mas os resultados ainda são insatisfatórios. Por esse motivo, esse trabalho tem como objetivo realizar uma investigacão a respeito do uso direto de sistemas atencionais como extratores de características locais. Para isso, é feita uma investigação em modelos computacionais de atenção e uma investigação realizada diretamente em mapas construídos a partir da resposta atencional humana, chamados de mapas de densidade. A investigação em modelos computacionais revelou que eles ainda não estão prontos para a tarefa de extração de características. No entanto, a investigação realizada diretamente com os dados da base de mapas de densidade, que foram construídos com o auxílio de

313 voluntários, revelaram que é possível e viável utilizar mapas de atenção visual como extratores locais de características. Os resultados dessa análise demonstraram que tanto qualitativamente quanto quantitativamente existe distinção e alta repetibilidade entre os pontos de fixação provenientes de humanos, mas que ainda precisam ser feitos inúmeros ajustes nos modelos computacionais, e na própria área de atenção visual computacional, para que a tarefa de extração de características seja realizável de forma eficiente pelos modelos computacionais de atenção.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A visão é o mais complexo dos sentidos do ser humano, é um processo que produz, a partir da visualização do mundo externo, uma descrição que é útil ao observador e que não é afetada por informações visuais irrelevantes do ambiente [9][10]. Quando estimulado, o sistema visual humano utiliza uma estratégia baseada na economia de esforço, dando prioridade à extração das informações necessárias à execução de uma determinada tarefa. Já entre os seres vivos que ocupam posições mais baixas da escala evolutiva, a percepção de estímulos provenientes do meio ambiente se constitui como atividade essencial no que diz respeito à sobrevivência [10].

Neste cenário, possui grande importância a chamada atenção visual, característica dos sistemas visuais biológicos responsável por selecionar as informações mais relevantes em um ambiente. Este mecanismo é determinante para a perpetuação e a evolução das espécies, na medida em que se caracteriza como a habilidade de fixar rapidamente a visão em pontos de interesse e reconhecer possíveis presas, predadores ou rivais [11].

Por ser uma tarefa de construção de descrições da realidade e de processamento de informação, a visão pode ser tratada computacionalmente [10]. Para realizar a percepção dos estímulos visuais em máquinas utiliza-se a visão computacional, que dispõe de um conjunto de métodos e técnicas para análise e interpretação de imagens.

No contexto da visão computacional, tem-se a atenção visual computacional (análoga à atenção visual biológica), dedicada a desenvolver mecanismos que reduzam o custo computacional no processamento de cenas visuais em máquinas. A partir de então, o termo atenção visual é utilizado para referenciar as versões computacional e biológica indistintamente.

É importante ressaltar que no meio científico, no que se refere à atenção visual biológica, apenas se possuem evidências de seu funcionamento. A partir destas evidências foram convencionados mecanismos para a realização da atenção visual computacional. Nesse sentido, acredita-se que o processamento visual se inicia a

partir de elementos simples e básicos, características discretas fornecidas pelos receptores visuais. Esses elementos se combinam perceptualmente por mecanismos involuntários do cérebro e do sistema visual para construir e formar padrões e formas identificáveis [12].

Do ponto de vista computacional, a atenção visual pode ser dividida em duas abordagens principais: a espacial (que se preocupa em determinar estaticamente quais regiões da cena são mais relevantes) e a temporal (que se preocupa em dinamicamente determinar em que região e momento do tempo realizar uma análise mais detalhada, ou seja, quais cenas e regiões nas cenas em uma sequência são as mais relevantes). A atenção pode ainda ser *bottom-up* (ou atenção baseada em características primitivas) ou *top-down* (baseada em modelos de objetos) [11].

Além desta classificação, os modelos de atenção ainda podem ser divididos quanto às diferentes técnicas utilizadas para detectar as áreas mais relevantes de uma cena. Assim, os modelos podem ser classificados entre cognitivos, bayesianos, gráficos, de análise espectral, baseados em teoria da informação, entre outros [13]. Estas técnicas apresentam como objetivo principal a construção de uma representação bidimensional na forma de um mapa, ponderando áreas mais informativas de um ambiente com valores altos e áreas menos relevantes com valores baixos. A essa medida de importância dá-se o nome de saliência e o conjunto de todas essas ponderações chama-se de mapa de saliência, também conhecido como o produto final da atenção visual em máquinas.

De um modo geral, a atenção visual tem sido bastante explorada no decorrer dos últimos anos. As pesquisas nessa área têm como objetivo principal produzir mapas de saliência que tenham uma alta similaridade com os mapas de densidade (ou mapas de calor), produzidos a partir de dados das fixações humanas coletados por meio de um equipamento de rastreamento ocular. As fixações humanas são pontos na imagem que representam a localização dos focos da retina das pessoas quando são solicitadas a analisar uma imagem. Esses pontos são utilizados para construir superfícies de densidade que indicam quais são as áreas que as pessoas olham com mais frequência. Essas superfícies são os mapas de densidade e são o padrão de referência para a área de atenção visual, de modo que todos os modelos computacionais são comparados com estes mapas para estimar a similaridade das abordagens com o sistema biológico atencional.

De acordo com os resultados obtidos em [14] alguns modelos de saliência já possuem resultados bastante significativos e próximos dos mapas de densidade. No entanto, ainda não existe uma vasta quantidade de aplicações na área de extração de características, utilizando os modelos de atenção computacionais, quando comparados a outras ferramentas clássicas já existentes, tais como a classe de algoritmos detectores e descritores de pontos-chave.

É comum encontrar na literatura a utilização de mapas de saliência como limitadores de espaço de busca para os algoritmos extratores de características. Um exemplo disso é a aplicação de localização de robôs móveis proposta por Siagian [2]. Essa aplicação utiliza o mapa de saliência clássico proposto por Laurent Itti [15], que tem como função ser uma ferramenta limitadora de regiões de busca para o algoritmo SIFT [16]. O algoritmo SIFT é responsável por identificar pontos-chave e realizar a correspondência dos pontos de uma região saliente com outra região saliente armazenada no banco de imagens do robô.

Segundo Siagian [2] o mapa de saliência foi utilizado apenas como ferramenta de pré-processamento de regiões devido à ausência de ferramentas biologicamente inspiradas que aproveitem diretamente as informações de saliência para a extração de características, já que a correta combinação do detector de pontos com o descritor de pontos é responsável pela alta taxa de correspondências entre duas imagens. Por conta dessa deficiência, os mapas de saliência não são amplamente utilizados em tarefas que exigem correspondência entre duas imagens, tarefas de reconhecimento ou até mesmo tarefas envolvendo odometria visual na robótica móvel. Para esse tipo de aplicação, normalmente os algoritmos detectores e descritores de pontos como o SIFT [16] e o SURF [17] são amplamente utilizados, mesmo possuindo um custo computacional elevado. Vale ressaltar que a atenção visual biológica tem um papel primordial nas etapas de recuperação de informação, reconhecimento de objetos, memória e aprendizado. No entanto, esse papel não tem sido totalmente refletido dentro da visão computacional.

Quando os mapas são utilizados como limitadores de região, a forma como essas regiões são selecionadas é crucial para o desempenho dos algoritmos extratores de características. Apesar dessa técnica apresentar uma redução do custo computacional do algoritmo detector e descritor, ela reduz bastante a quantidade de pontos encontrada pelos detectores, já que eles são mais apropriados para funcionar em regiões grandes. Essa redução fragiliza uma aplicação de rastreamento que precisa confiar em uma quantidade muito reduzida de pontos, sendo que alguns deles podem ser *outliers*. Além de fragilizar a etapa de rastreamento, é preciso segmentar e selecionar as regiões de forma bastante cuidadosa, como explicado por Siagian em seu trabalho [2]. Segundo ele, a forma de segmentar e selecionar as regiões precisa garantir um preenchimento de pelo menos 60% da cena para resultados significativos. O próprio Siagian utiliza um método iterativo de segmentação e faz várias restrições ao tamanho das regiões, gerando custos adicionais ao método proposto por ele.

Por outro lado, os mapas de saliência quanto utilizados como extratores de áreas de interesse possuem alguns aspectos interessantes com relação a algoritmos clássicos, o primeiro aspecto é a velocidade de processamento, já que eles selecionam

rapidamente apenas as porções mais informativas do ambiente. O segundo aspecto é que os mapas de saliência não são dependentes de contexto ou de apenas uma característica do ambiente, como o SIFT e o SURF por exemplo, que são extremamente dependentes da presença de bordas abruptas na cena para encontrarem pontos considerados relevantes. Além disso, mapas de saliência são inspirados no processo atencional humano, que é responsável pela seleção de estímulos que serão utilizados no reconhecimento de objetos, na recuperação de informação, no próprio aprendizado e até mesmo nas ações motoras realizadas pelo ser humano. Então, produzir sistemas computacionais de extração de características que tenham como inspiração o melhor aproveitamento possível das informações provenientes de um mapa de saliência é de crucial importância para o desenvolvimento de sistemas de visão computacional.

Nesse sentido, nota-se que existe potencial para aplicações envolvendo mapas de saliência diretamente como extratores de pontos de interesse. No entanto, não há estudos na literatura que demonstrem essa possibilidade. Assim, esse trabalho tem como objetivo investigar a possibilidade de utilizar sistemas de atenção visual como ferramenta de detecção e descrição de pontos de interesse em cenas.

1.1 Objetivos

1.1.1 Objetivos gerais

O objetivo deste trabalho é investigar a possibilidade de utilizar sistemas atencionais não somente como limitadores de espaço de busca, e sim diretamente na tarefa de detecção e descrição de características.

1.1.2 Objetivos específicos

Para atingir os objetivos deste trabalho, são propostos os seguintes objetivos específicos:

- Realizar uma revisão bibliográfica sobre os modelos computacionais de atenção visual existentes, assim como uma revisão a respeito das bases de mapas de densidade de fixação existentes;
- Realizar uma revisão bibliográfica sobre os detectores de pontos-chave e sobre os descritores clássicos encontrados na literatura;
- Selecionar alguns modelos computacionais de atenção que estão disponíveis e verificar se são capazes de realizar a tarefa de detecção de pontos de interesse;
- Se os modelos de atenção computacionais forem capazes de realizar essa tarefa, realizar uma análise dos descritores existentes e verificar se podem ser utilizados para descrever pontos proveninentes de mapas de saliência;
- Avaliar as vantagens e desvantagens da utilização de descritores clássicos utilizados para descrever pontos de saliência e propor melhorias;
- Avaliar as melhorias propostas realizando testes comparativos utilizando sequências de imagens;
- Se os modelos de atenção computacional ainda não forem capazes de realizar a tarefa de detecção de pontos interesse, realizar a investigação em mapas de densidade;
- A partir dos resultados obtidos através de análises em mapas de densidade concluir se há possibilidade de utilizar modelos atencionais como extratores de características locais, e se podem ser realizadas modificações nos modelos computacionais para realizar tal tarefa.

1.2 Estrutura do Trabalho

O trabalho a seguir se divide em capítulos da seguinte maneira:

Capítulo 2 - apresenta uma revisão bibliográfica a respeito da atenção visual biológica, dos modelos computacionais de atenção existentes, das bases de mapas de fixação encontradas na literatura e seus aspectos de construção, assim como uma revisão dos principais detectores e descritores de características;

Capítulo 3 - são discutidos os trabalhos que fazem uso dos mapas de saliência como ferramentas de pré-processamento, mostrando desvantagens e vantagens de cada abordagem;

Capítulo 4 - apresenta as investigações realizadas em modelos computacionais de atenção e os resultados obtidos;

Capítulo 5 - apresenta as investigações realizadas em mapas de fixação e os resultados obtidos;

Capítulo 6 - são apresentadas as considerações finais e os trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

Neste capítulo será apresentada uma breve revisão bibliográfica necessária ao desenvolvimento deste trabalho. Ela será iniciada com a Seção 2.1, que apresenta uma explanação a respeito da inspiração biológica da atenção visual, seguida pela Seção 2.2 com um aprofundamento referente aos modelos computacionais de atenção. Na Seção 2.3, serão apresentadas as principais bases de mapas de densidade e fixação encontradas na literatura e que são referência para o desenvolvimento da área. Por fim, na Seção 2.4 são apresentados os principais detectores e descritores de características locais encontrados na literatura.

2.1 Inspiração Biológica da Atenção Visual

A todo instante, os olhos se deparam com uma enorme carga de estímulos visuais. No entanto, é impossível processar toda a informação que chega aos olhos de uma só vez. Diante disto, o sistema visual se comporta como um sistema sensorial somático, onde existem camadas envolvidas com o processamento de diferentes aspectos das informações visuais [18].

Segundo Itti, Koch [11] e Miller [19], a informação visual é inicialmente captada pelos fotorreceptores da retina e, de forma organizada, é conduzida para as outras áreas do cérebro, como o córtex visual primário (V1) e as áreas extra-estriadas chamadas V2, V3, V4 e V5, conhecidas como um conjunto de áreas de ordem superior que também recebem informações a partir da retina e do córtex visual primário.

Depois que os impulsos nervosos abandonam a retina, dirigem-se através dos nervos ópticos onde encontram o núcleo geniculado lateral, que é uma área responsável pelo controle da quantidade de sinais que tem permissão para chegar até o córtex visual. Quando estes impulsos chegam ao córtex visual, eles são recebidos por vias paralelas de milhões de colunas verticais de células neuronais, onde cada coluna representa uma unidade funcional.

A Figura 2.1 ilustra algumas destas vias, em que algumas são especializadas

em identificar as posições tridimensionais dos objetos no espaço, enquanto outras analisam a cor, a movimentação e a forma dos elementos de uma cena. Além disso, é importante notar a interação entre as camadas V1, V2, V3, V4 e V5, ilustrando o fluxo da informação e como a atenção parte da integração de estímulos simples do ambiente, de modo que as camadas superiores são responsáveis pelas funções mais complexas da atenção visual.

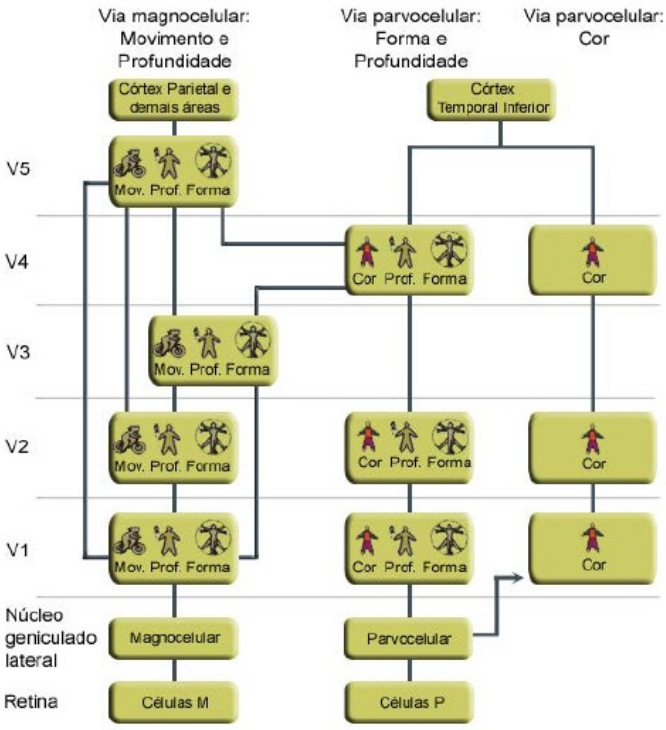


Figura 2.1: Vias paralelas do sistema visual humano [1].

A partir do córtex visual primário, o fluxo da informação prossegue por dois caminhos paralelos, o primeiro passando por áreas corticais incluindo o córtex parietal, responsável principalmente pela localização espacial e por direcionar a atenção para objetos de interesse em uma cena, e o segundo pelo córtex temporal inferior, responsável principalmente pelo reconhecimento e identificação de estímulos visuais [11]. Ou seja, o córtex temporal inferior se preocupa com *o que* está sendo visto de um modo geral, enquanto que o córtex parietal se interessa mais em encontrar *onde* estão os objetos.

Estes dois mecanismos atuam de forma paralela a partir da informação proveniente do córtex visual, sendo que a atenção passa a ser tanto direcionada involuntariamente por estímulos que emergem do ambiente, quanto voluntariamente para alvos específicos, de acordo com a importância atribuída pelo observador.

Para um comportamento coerente, estas duas formas de atenção atuam juntas. Para realizar essa integração, estes dois tipos de estímulos chegam ao córtex pré-frontal, cujo fluxo é ilustrado na figura 2.2. De acordo com Miller [19], o córtex

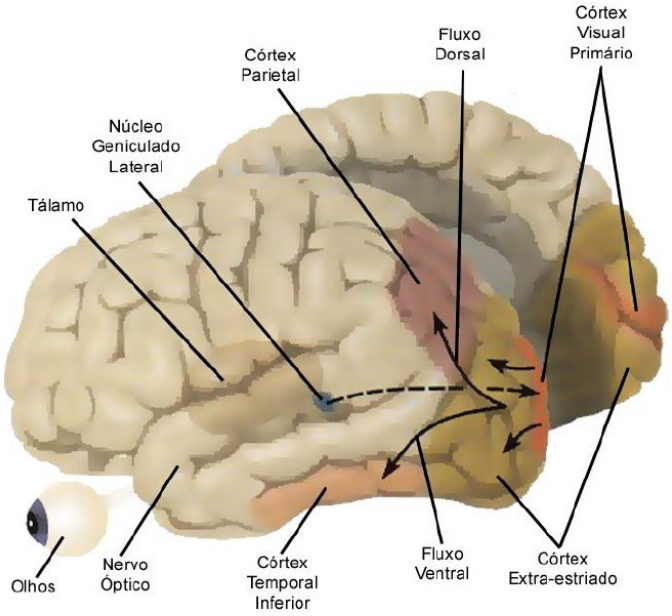


Figura 2.2: Fluxo da informação através do córtex visual [1].

pré-frontal é o centro de controle da atenção visual, pois ele é responsável tanto pelo planejamento da ação, quanto pela modulação da atenção envolvida em comportamentos relacionados ao reconhecimento, planejamento e controle motor de um indivíduo.

Além do caminho entre a retina e o córtex visual, os sinais visuais também fluem para outras áreas do cérebro como o colículo superior, que é uma região também importante para a atenção visual, justamente por controlar os movimentos oculares que fazem com que os olhos se posicionem sobre uma porção discreta do campo visual. A principal razão biológica para a existência desses movimentos é a necessidade de localizar a fóvea nas regiões mais informativas em um ambiente, tendo em vista que, na retina, a fóvea é a região em que há maior densidade de fotorreceptores [20].

Com ajuda destes mecanismos trabalhando em paralelo, o ser humano consegue perceber, numa fração de tempo reduzida, se existe algo peculiar em uma cena, de modo que os olhos se fixam em um ponto importante no campo visual, saltando de um local da cena obsrvada para o outro duas ou três vezes por segundo. Esses movimentos que proporcionam um varredura rápida em todo o ambiente também são chamados de micro-sacadas, e são fundamentais para a **extração de todas as características relevantes de uma cena**.

Diante dos indícios do funcionamento do sistema visual humano, diversas áreas da ciência visam compreender melhor este mecanismo, cuja principal motivação está no fato de que seres humanos podem reagir não somente reflexivamente diante das informações sensoriais imediatas e importantes, mas também pela possibilidade de substituir ou aumentar reações habituais, modulando o comportamento de acordo

com intenções próprias. Segundo Corbetta [21] e Fintrop [22] a atenção é quem define essa capacidade mental para selecionar estímulos, respostas, memórias ou pensamentos que são comportamentalmente relevantes.

Neste sentido, ainda há muitas questões em aberto, como por exemplo, vários neurocientistas têm se preocupado em analisar o que ocorre no cérebro quanto à integração de vários estímulos visuais, e como o controle da atenção é realizado de fato. Os questionamentos e as ideias desta área da neurociência também despertam o interesse dos pesquisadores de visão computacional, que tentam reproduzir este mecanismo na forma de sistemas de visão que sejam capazes de processar e identificar os pontos mais importantes do ambiente em tempo real. Sendo assim, a próxima seção deste trabalho detalha os principais modelos computacionais de atenção visual que surgiram no decorrer das duas últimas décadas, inspirados no modelo de atenção visual humano, e que são utilizados nesse trabalho.

2.2 Modelos Computacionais da Atenção Visual

Para fundamentar as descobertas encontradas pela neurociência à respeito da atenção visual, foi proposta a teoria de integração de características *(Feature Integration Theory)* por Treisman e Gelade [23]. Esta teoria propõe que atributos simples do campo visual como cor, intensidade e orientação são registrados automaticamente e de forma paralela através de todo o campo visual, enquanto que os objetos são identificados separadamente, em um estágio posterior, o que exige o direcionamento da atenção.

Itti e Koch foram os primeiros a desenvolver computacionalmente um modelo de atenção visual baseado na teoria de integração de Treisman e Gelade. Desde então, durante as duas últimas décadas, o interesse por esta área se intensificou e surgiram diversas abordagens de modelos de atenção. Assim, uma classificação dos diferentes tipos de mapas de saliência é inevitável.

Atualmente, existem alguns critérios que são considerados para classificar os modelos baseados em mapas de saliência. Uma forma de classificar os modelos está entre **dinâmicos** e **estáticos**. Nos modelos estáticos os métodos se preocupam apenas em determinar estaticamente quais regiões da cena são mais relevantes para uma análise posterior. Já nos modelos dinâmicos, a seleção visual é dependente tanto da saliência atual quanto do conhecimento acumulado de pontos.

Além desta classificação, também é comum dividir os modelos de acordo com os critérios utilizados para construir o mapa de saliência. De acordo com esta nomenclatura, os modelos podem ser divididos entre: cognitivos, bayesianos, que utilizam a teoria da informação, que utilizam grafos, modelos no domínio da frequência e modelos que utilizam ferramentas de aprendizado de máquina. Mais detalhes sobre

esses modelos podem ser encontrados em [13].

No que diz respeito à similaridade com os mapas de atenção provenientes de fixações humanas, os modelos que utilizam técnicas de Deep Learning apresentam resultados de similaridade melhores e hoje são tidos como as abordagens mais próximas dos mapas de densidade. Por esse motivo, se densenvolveram diversos modelos utilizando essas técnicas e todos com resultados muito semelhantes. Por conta da alta similaridade entre as abordagens que utilizam Deep Learning, foram escolhidos 2 modelos que apresentam a maior similaridade já obtida com os mapas de densidade e representam bem essa categoria de mapas. Esses modelos são o Sam-VGG [5] e o Sam-ResNet [5] (Seção 2.2.4).

Há alguns modelos que têm uma similaridade significativa com os mapas de densidade, porém não tão elevada quanto os modelos que utilizam *Deep Learning*. Mesmo utilizando técnicas mais simples, eles apresentam um grau de similaridade significativo com os mapas de densidade. Para representar essa categoria de mapas foram escolhidos dois representantes que apresentam maior similaridade com os mapas de densidade, que são o LDS [4] e o GBVS [24] (Seção 2.2.3 e 2.2.2, respectivamente).

Por fim, foi escolhido também o modelo clássico proposto por Itti [3] na investigacão, pois ele apesar de estar mais distante no *ranking* de similaridade com os mapas de densidade, foi o primeiro mapa de saliência criado na literatura (Seção 2.2.1). Além disso, apesar de apresentar uma similaridade baixa, ele possui aspectos mais próximos de um mapa de saliência que de um detector de bordas, já que alguns mapas de saliência posteriores muitas vezes são mais distantes da resposta de saliência e mais próximos de detectores de borda que a abordagem pioneira da área.

2.2.1 Modelo Clássico de Itti e Koch

O modelo de atenção visual desenvolvido por Laurent Itti e Christof Koch é um método clássico de atenção visual, classificado como um modelo cognitivo e amplamente estudado por muitos pesquisadores.

Este modelo consiste na decomposição da imagem em três características primitivas: cor, intensidade e orientação. Para cada uma delas são criados diferentes mapas. As diferentes regiões desses mapas são qualificadas segundo a resposta gerada pelas características e as regiões que mais se destacam são chamadas de áreas de atenção. Os mapas são então combinados para a criação do mapa de saliência final, definindo os focos da atenção. Para isso, este processo é seguido por algumas etapas, como ilustrado na Figura 2.3.

No processo de extração de características são criados mapas com três características primitivas onde as regiões claras representam locais que possuem abundância

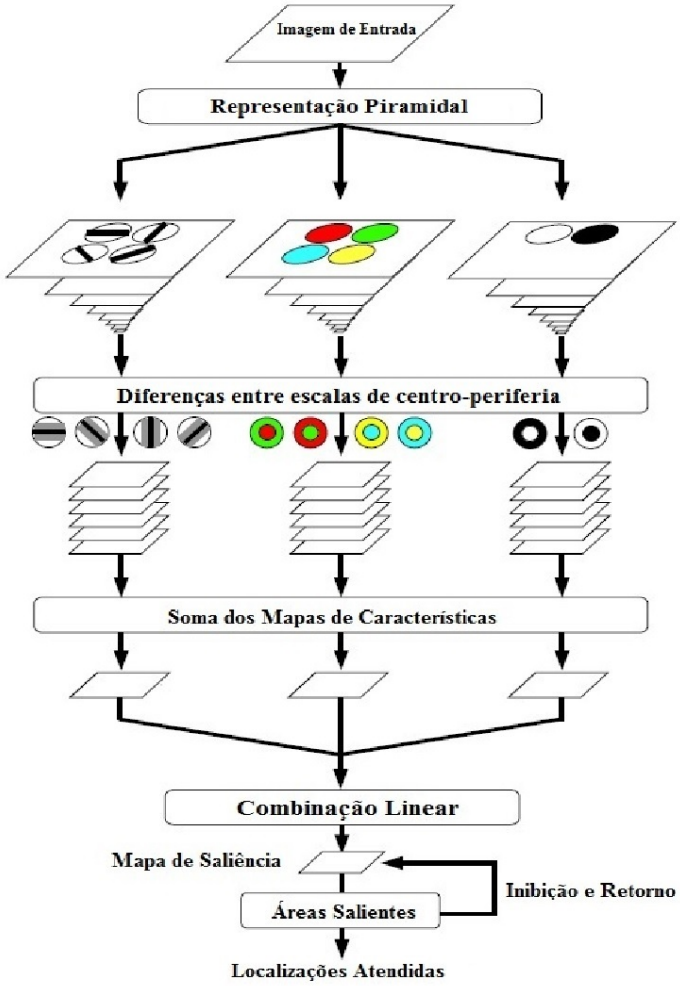


Figura 2.3: Etapas de construção do mapa de saliência clássico [2].

de uma determinada característica, e escuras para os locais que possuem a ausência dela. Nessa etapa, são gerados nove mapas (1 mapa de intensidade, 1 mapa vermelho, 1 mapa amarelo, 1 mapa azul, 1 mapa verde e 4 mapas de orientação).

Os mapas de orientação são gerados através de convoluções com filtros de gabor [25] em ângulos de 0°, 45°, 90° e 135°. Para cada mapa formado são criados 9 mapas em escalas distintas para cada característica, de modo que a escala 1 corresponde ao mapa original e a escala 9 corresponde ao mapa com a menor resolução, formando assim pirâmides gaussianas. Pirâmides gaussianas são representações multi-escala de uma determinada imagem, considerando que o nível mais baixo da pirâmide seja a imagem original e cada nível superior seja uma amostragem suavizada da imagem do nível imediatamente abaixo. Com as pirâmides definidas, são formados 81 mapas, obtidos efetuando operações de diferença entre os mapas com escalas distintas. Este processo, segundo Itti et. al [15] [3], tenta representar a ação dos neurônios dos campos visuais, que são estimulados em uma pequena região do espaço

visual central e são inibidos em áreas vizinhas.

A operação de diferenças entre mapas em escalas distintas é realizada da seguinte forma:

- Para o canal de intensidade são gerados 6 mapas resultantes da operação de subtração ponto-a-ponto dos pares de mapas nas escalas (2-5), (2-6), (3-6), (3-7), (4-7), (4-8);

- Para o canal de cor são gerados 12 mapas utilizando a teoria de oponência das cores. Sendo que 6 mapas são resultantes de duas operações de subtração dos mapas vermelhos pelos mapas verdes nas escalas *[(2 (vermelho) - 2 (verde)) - (5 (verde) - 5 (vermelho))], [(2 (vermelho) - 2 (verde)) - (6 (verde) - 6 (vermelho))], [(3 (vermelho) - 3 (verde)) - (6 (verde) - 6 (vermelho))], [(3 (vermelho) - 3 (verde)) - (7 (verde) - 7 (vermelho))], [(4 (vermelho) - 4 (verde)) - (7 (verde) - 7 (vermelho))], [(4 (vermelho) - 4 (verde)) - (8 (verde) - 8 (vermelho))]*;

- Mais 6 mapas resultantes de duas operações de subtração dos mapas azuis pelos mapas amarelos nas escalas *[(2 (azul) - 2 (amarelo)) - (5 (amarelo) - 5 (azul))], [(2 (azul) - 2 (amarelo)) - (6 (amarelo) - 6 (azul))], [(3 (azul) - 3 (amarelo)) - (6 (amarelo) - 6 (azul))], [(3 (azul) - 3 (amarelo)) - (7 (amarelo) - 7 (azul))], [(4 (azul) - 4 (amarelo)) - (7 (amarelo) - 7 (azul))], [(4 (azul) - 4 (amarelo)) - (8 (amarelo) - 8 (azul))]*;

- Para o canal de orientação são gerados 24 mapas. Sendo que para cada orientação do filtro de gabor é realizada a subtração ponto-a-ponto dos pares de mapas nas escalas *(2-5), (2-6), (3-6), (3-7), (4-7), (4-8).*

No total, 42 mapas são computados, 6 de intensidade, 12 de cor e 24 de orientação, pois são construídos 6 mapas para cada uma das quatro orientações utilizadas.

Para combinar os mapas de características em somente um é necessário utilizar um operador de normalização *N(.)*. Neste modelo de atenção existem quatro métodos para realizar a normalização, quais sejam (i) Somatório Simples, (ii) Normalização não linear, (iii) Competição Iterativa, (iv) Normalização com pesos aprendidos. Em [26] [27] é feita uma comparação entre os resultados de mapas de saliência obtidos utilizando cada operador de normalização e, por conta da simplicidade e dos resultados satisfatórios encontrados, é mais comum utilizar nas aplicações a competição iterativa, que consiste na aplicação de filtros 2D de diferenças gaussianas (DoG).

Uma vez escolhida a estratégia de normalização que melhor se adapte aos objetivos de utilização do mapa de saliência, os mapas de características são então

normalizados e combinados em três mapas de conspicuidade (I, C e O), gerados através de uma adição entre escalas, que consiste em expandir cada mapa para a escala 4 e efetuar uma soma pixel a pixel.

Para compor o mapa de saliência os três mapas (I, C e O) são normalizados e uma combinação linear é realizada entre eles, o que produz a saída final S, como ilustrado na Figura 2.4.

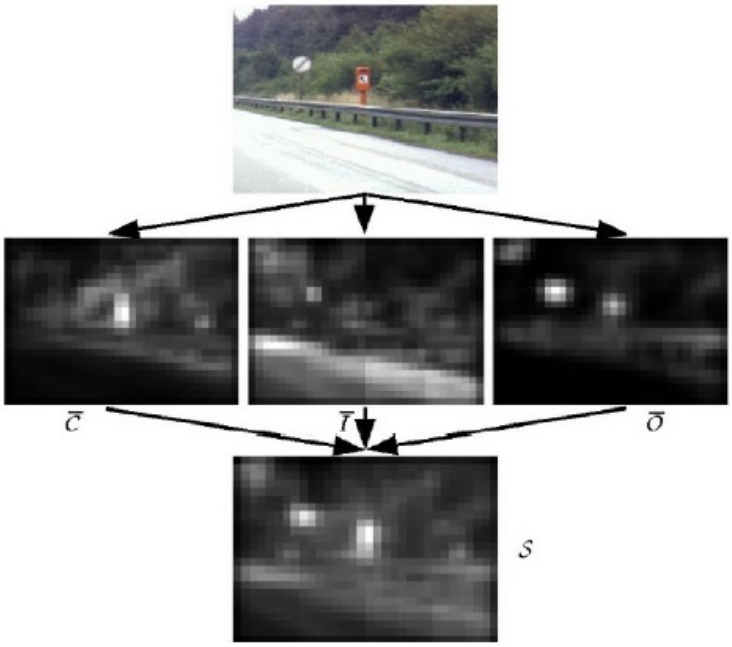


Figura 2.4: Mapa de saliência S [3].

Junto com o modelo de atenção, Itti também propôs uma técnica para selecionar os focos atencionais. Essa técnica tem como objetivo simular o movimento de sacadas dos olhos humanos. Essa técnica é composta por neurônios do tipo Integra e Dispara [15] [1], que têm como função representar o mapa de saliência S, de modo que a estimulação externa de cada neurônio é definida pelo valor de saliência dos respectivos pontos no mapa de saliência.

Por sua vez, a rede de neurônios Integra e Dispara alimenta uma rede neural do tipo WTA (Winner-Takes-All) [15] [28], considerada uma arquitetura neural plausível para descobrir a localização mais saliente no mapa de saliência, uma vez que é capaz de determinar um ponto de interesse representado por um neurônio vencedor. É importante notar que os neurônios da rede Integra e Dispara são utilizados, neste caso, somente como integradores dos valores de S, onde cada neurônio tem como função ativar seu correspondente neurônio WTA.

Na WTA, todos os neurônios recebem ativação de forma independente, até que o neurônio vencedor *(winner)* alcança o limiar e dispara, desencadeando simultaneamente três mecanismos: primeiro, o foco da atenção é direcionado para a localização do neurônio vencedor; segundo, o inibidor global da WTA é acionado e todos os de-

mais neurônios da WTA são inibidos; terceiro, a região sob o foco da atenção de tamanho fixo é temporariamente inibida na rede de neurônios Integra e Dispara, permitindo que a próxima região saliente seja destacada, garantindo também que o foco da atenção não seja novamente direcionado para a região anterior, caracterizado um mecanismo de inibição de retorno (mais detalhes podem ser encontrados em [15]). Uma representação visual do funcionamento dessa técnica é ilustrada na Figura 2.5.

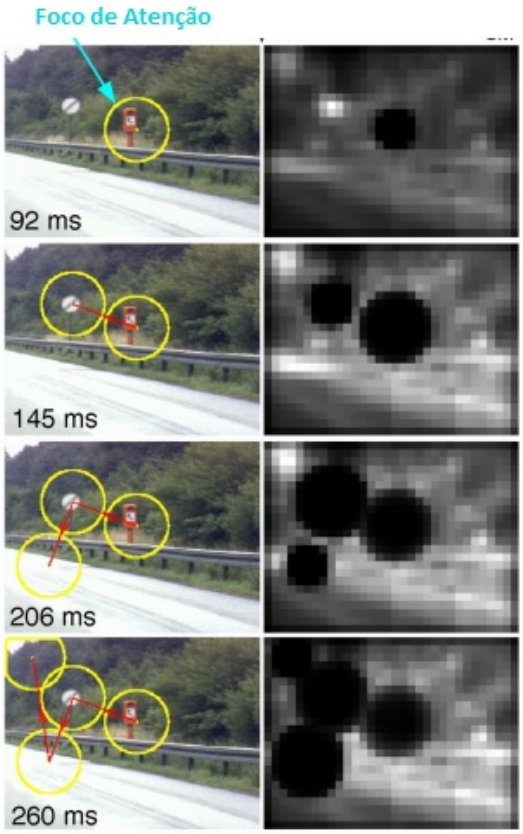


Figura 2.5: Representação visual da rede *winner-takes-all* funcionando com 4 iterações [3].

2.2.2 GBVS

No modelo GBVS os mapas são extraídos de forma semelhante à apresentada pelo modelo clássico de Itti e Koch. Com a diferença de que para cada mapa de característica é gerado um grafo totalmente conectado, de modo que nodos são responsáveis pela representação de todas as localizações dos pontos. Pesos entre os nós são associados proporcionamente à similaridade em relação às suas características e pesados por suas distâncias espaciais. Em um processo de normalização, nós que representam altos valores de dissimilaridade em relação à vizinhança são definidos com altos valores de saliência. Os mapas são finalmente normalizados para enfatizar

(a) (b)

Figura 2.6: Mapa de saliência GBVS. (a) Imagem de entrada. (b) Mapa GBVS.

os detalhes importantes, e por fim combinados em um único mapa de saliência. A Figura 2.6 ilustra um exemplo de mapa de saliência GBVS.

2.2.3 LDS

O modelo de atenção LDS é um modelo de atenção visual que tem como objetivo distinguir os alvos e distrações que compartilham certos atributos visuais [4].

Para isso, este modelo utiliza um grande conjunto de imagens de treinamento. Junto às imagens de treinamento também são utilizados os mapas de fixação humana disponibilizados publicamente por diversos pesquisadores em http://saliency.mit.edu/. Esses mapas definem os alvos corretos para a cena.

Inicialmente são gerados diferentes mapas de características e, com o auxílio da técnica PCA [29][30], os elementos salientes de cada mapa são ressaltados. Esses mapas são gerados em diferentes escalas. Dispondo do mapa de fixação da imagem de entrada e dos mapas de características para a imagem, os mapas de características são separados em cinco grupos distintos: (i) mapas que ressaltam apenas os elementos salientes, (ii) mapas que ressaltam apenas os distratores, (iii) mapas que ressaltam tanto distratores como elementos salientes, (iv) mapas que apresentam respostas baixas e elevadas para partes diferentes dos elementos salientes, (v) mapas que não conseguem distinguir quem são os alvos e os distratores. Esse último grupo é eliminado da fase de treinamento [4]. Posteriormente, é aplicado um fator de normalização nos mapas de características escolhidos para elevar o contraste entre alvos e distratores.

Em seguida, os mapas de características e a imagem são quebrados em *patches* e com a informação do mapa de fixação humano é possível separar os *patches* da imagem que indicam os alvos e os *patches* que indicam os distratores. Cada *patch* da imagem possui n vetores de características, que são resultantes dos n mapas criados. A partir de então, esses vetores são utilizados em um algoritmo de treinamento que gera como saída uma escolha otimizada de características e pesos para cada mapa escolhido; a junção dos pesos com os mapas de características forma o mapa de

saliência final. A Figura 2.7 ilustra a fase de treinamento e a fase de teste da abordagem LDS.

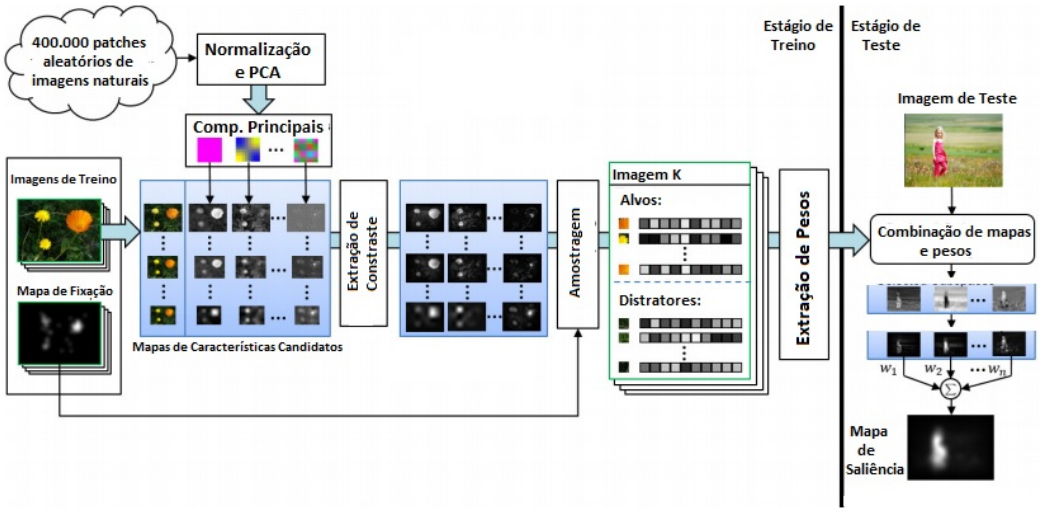


Figura 2.7: Etapas de construção do mapa de saliência LDS [4].

2.2.4 SAM-VGG e SAM-ResNET

Os métodos tradicionais de predição de saliência seguiram as evidências biológicas definindo características que capturam pistas de baixo nível, como cor, contraste e textura ou conceitos semânticos como rostos, pessoas e texto. No entanto, essas técnicas não conseguiram capturar a ampla variedade de causas que contribuem para a definição de mapas de saliência visual. Com o advento das redes neurais profundas, a previsão de saliência alcançou fortes melhorias, graças a arquiteturas específicas e a grandes conjuntos de dados.

A arquitetura de atenção denominada SAM (Figura 2.8) inicialmente calcula uma grande quantidade de mapas de características da imagem de entrada através de uma arquitetura de redes convolucionais (CNN) [31]. Em seguida, o conjunto de características extraídas da imagem de entrada alimenta um módulo atencional composto por uma rede LSTM [31] modificada, chamado de Atentive CONVLSTM. Esse módulo é utilizado para construir o mapa de saliência. Ele possui uma LSTM estendida e modificada para trabalhar com características espaciais, já que originalmente as LSTMs não são diretamente empregadas para predição de saliência, pois trabalham com sequências de vetores variando no tempo. Isso foi conseguido substituindo algumas operações por convoluções nas equações da LSTM. Além disso, a natureza sequencial desse tipo de rede foi explorada de maneira iterativa em vez de usar o modelo para lidar com depêndencias temporais.

Deste modo, o módulo atencional é responsável por realizar modificações progressivas do mapa de saliência, ajustando seus focos e elimando os elementos falsamente

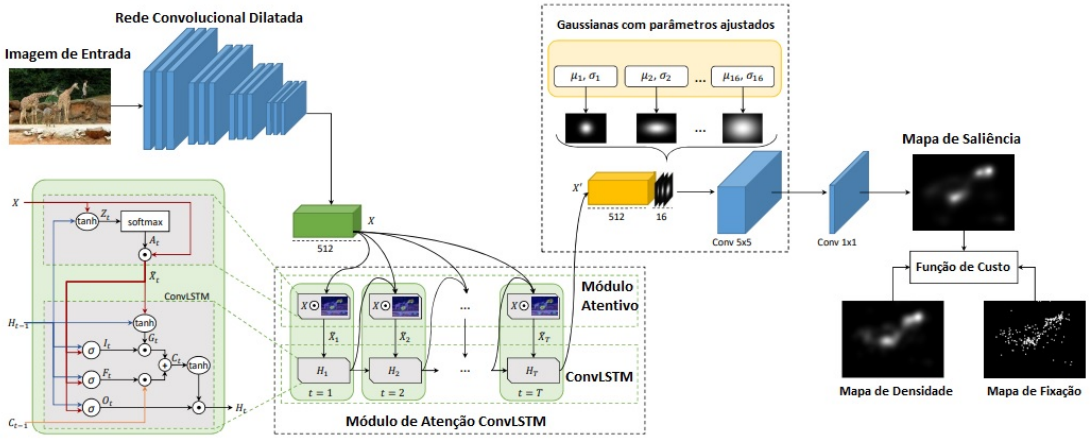


Figura 2.8: Modelo de Atenção SAM (Saliency Attention Model) [5].

identificados como salientes, como ilustrado pela Figura 2.9.

As saídas do módulo atencional são convolvidas com 16 gaussianas, cujos valores de média e desvio padrão são livremente aprendidos durante a fase de treinamento. A idéia de convolver os dados com essas funções procura melhorar a similaridade com os mapas de densidade, pelo fato deles serem formados por funções gaussianas. Ao final desse processo o mapa de saliência está criado.

Para treinar o modelo, é utilizada uma função de custo. Essa função foi definida como a combinação ponderada de 3 métricas que são utilizadas para comparar mapas de saliência com mapas de densidade, sendo elas: NSS *(Normalized Scanpath Saliency)* [32], Coeficiente de Correlação de Pearson [32] e Divergência de Kullback-Leibler [32].

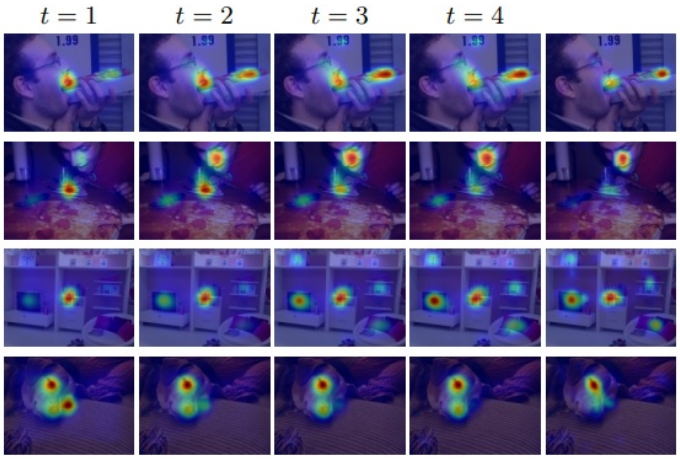


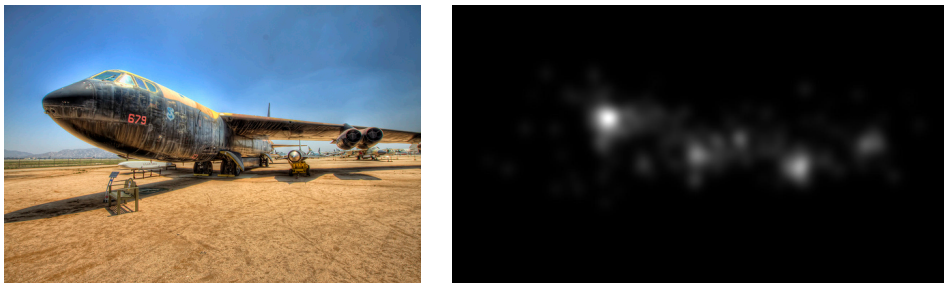
Figura 2.9: Etapas de refinamento dos focos atencionais realizado pelo módulo ConvLSTM do modelo SAM [5].

Essa arquitetura atencional definida como SAM originou duas abordagens: SAM-ResNet e SAM-VGG. Basicamente a diferença entre elas está apenas na CNN utilizada na primeira etapa para extrair as características da imagem. Na SAM-ResNet

utiliza a rede convolucional ResNet [5]. Ela consiste de 5 blocos convolucionais e uma camada final totalmente conectada. Os blocos são compostos por redes convolucionais que reduzem a dimensão dos mapas por passos de 2. Já na abordagem SAM-VGG é utilizada a rede VGG [5], composta por 13 camadas convolucionais e 3 camadas totalmente conectadas. As camadas convolucionais são divididas em 5 blocos convolucionais, onde cada uma delas é seguida por uma camada de reescalonamento com passo igual a 2.

2.3 Mapas de Fixacão e Mapas de Densidade

Mapas de fixação são mapas de atenção visual construídos através do rastreamento ocular das fixações humanas, realizado por um equipamento denominado *eye-tracking*. Utilizando esse equipamento, pessoas são solicitadas a visualizar diferentes tipos de cenas por alguns segundos, enquanto seu rastreamento ocular é mapeado. Ao final do processo de coleta de dados, há várias fixações com suas respectivas posições *(x,y)* no plano da imagem. O conjunto de todas as fixacões dispostas no plano da imagem compõe o mapa de fixação. A partir do mapa de fixação são construídos os mapas de densidade. Para criá-los é atribuída uma função gaussiana sobre cada ponto, produzindo ao final do processo uma superfície (Figura 2.10). Essa superfície é utilizada como referência para comparar os modelos de atenção computacional criados com o sistema atencional biológico.



(a) (b)

Figura 2.10: Mapa de densidade das fixações. (a) Imagem original. (b) Mapa de densidade.

Na literatura já existem diversas bases de mapas produzidas dessa forma, publicadas e algumas disponibilizadas para seu uso em pesquisas. De um modo geral existem dois tipos de bases disponíveis: as bases de imagens, para testes utilizando modelos de atenção estática, e as bases de vídeo utilizados para testes em modelos de atenção dinâmicos.

O que foi observado é que cada base foi construída com um propósito específico. As Tabelas 2.1 a 2.3 apresentam de forma resumida os objetivos das bases estáticas

e dinâmicas que foram encontradas na literatura.

Base de Dados	Objetivo	Voluntários	Tempo (s)	Equipamento	Base Disponível
FiFA	Demonstrar que as faces das pessoas atraem atenção significativa.	8	2	Eyelink 1000 (1000Hz)	Não
IRCCyN Image 1	Demonstrar o comportamento da atenção em cenas coloridas de imagens naturais com variadas resoluções.	40	15	Cambridge Research (50 Hz)	Não
IRCCyN Image 2	Verificar a capacidade do ser humano de prever os principais objetos da cena.	18	15	Cambridge Research (50 Hz)	Não
KTH	Validar um modelo computacional e concluir sobre as fixações humanas em áreas simétricas da imagem.	31	5	Eyelink I	Não
LIVE DOVES	Verificar o comportamento da atenção utilizando cenas sem objetos facilmente identificados como interessantes.	29	5	Fourward Tech. Gen. V (200 Hz)	Não
MIT Benchmark	Disponibilizar para a comunidade acadêmica uma base de 300 imagens.	39	3	ETL 400 ISCAN (240Hz)	Sim
MIT CSAIL	Disponibilizar para a comunidade acadêmica cenas de imagens naturais.	15	3	-	Sim
MIT CVCL	Entender os padrões do movimento ocular orientados por tarefa.	14	-	ISCAN RK-464 (240 Hz)	Sim

Tabela 2.1: Bases de mapas de densidade existentes utilizando imagens estáticas.

Base de Dados	Objetivo	Voluntários	Tempo (s)	Equipamento	Base Disponível
MIT LowRes	**Verificar como a resolução da imagem afeta a consistência das fixações oculares.**	**64 (8 por img)**	**3**	**ETL 400 ISCAN (240 Hz)**	**Sim**
NUSEF	Estudar padrões de visualização em imagens semanticamente ricas e diversificadas incluindo rostos, retratos, cenas internas/externas e conteúdo afetivo.	25	5	ASL	Não
Toronto	Extrair gravações de movimento dos olhos usando cenas naturais para validar um modelo de saliência.	20	4	Câmera Hitachi CCD	Não
TUD Interactions	Estudar a influência da atenção na construção de métricas para avaliar a qualidade de uma imagem.	20	10	iView X RED (50Hz)	Não
CAT2000	Disponibilizar para a comunidade acadêmica uma base de 2000 imagens.	120 (24 por img)	5	EyeLink1000 (1000Hz)	Sim
EMOd	Estudar a influência dos sentimentos na atenção visual.	16	3	Eyelink 1000 (1000Hz)	Sim
FiWI	Estudar a atenção visual em páginas web.	11	5	Eyelink 1000 (1000Hz)	Não
FIGRIM	Estudar os efeitos da memória na atenção visual.	15	2	Eyelink 1000 (500Hz)	Não
Ehinger	Observar os processo de buscar pessoas em variadas cenas.	14	-	ISCAN RK-464 (240Hz)	Não
VIP	Observar a atenção em meio a imagens neutras e afetivas.	75	5	SMI RED 250 (120Hz)	Não

Tabela 2.2: Bases de mapas de densidade existentes utilizando imagens estáticas.

Base de Dados	Objetivo	Voluntários	Tempo (s)	Equipamento	Base Disponível
Actions	Modelar os movimentos oculares utilizando cenas de filmes e lutas esportivas.	16	60	SMI iView X HiSpeed	Não
ASCMN	Testar os movimentos oculares em vídeos do tipo vigilância caracterizados por objetos em movimento.	13	2-76	faceLAB	Não
DIEM	Mostrar a influência que elementos em movimento na cena exercem sobre as fixações humanas.	42	27-217	Eyelink 2000	Sim
GazeCom Video	Estudar a variabilidade dos padrões do movimento dos olhos durante a visualizacão de diferentes cenas naturais e psicofísicas.	54	20	EyeLink II	Não
USC CRCNS Orig	Estudar o papel de fatores como a memória na atenção visual em cenas dinâmicas.	8	6-90	ISCAN RK-464	Não
USC CRCNS MTV	Estudar o papel de fatores como a memória na atenção visual em clipes com cenas abruptas.	16	1-3	ISCAN RK-464	Não
EyeTrackUAV	Monitorar o comportamento humano observando veículos aéreos.	14	14:47	EyeLink 1000 Plus (1000 Hz)	Não

Tabela 2.3: Bases de mapas de fixação existentes utilizando sequências de vídeos.

Como apresentado nas Tabelas 2.1 a 2.3 algumas bases não estão disponíveis para uso, e as que estão não suprem as necessidades de análise das fixações humanas como detectores de características, pois não são utilizadas imagens com várias transformações, como por exemplo, imagens rotacionadas, imagens com mudanças de escala, imagens comprimidas ou com mudanças de iluminação. As bases existentes e disponíveis estão mais preocupadas em verificar outros aspectos da atenção humana, como atenção em ambientes dinâmicos, em ambientes neutros ou com algum ponto específico a ser analisado.

Apenas a base MIT LowRes apresenta uma base com imagens borradas. No entanto, os testes foram realizados apenas com o borramento e o objetivo não era discutir aspectos de invariância dos pontos de fixação em transformações, mas sim avaliar se seria possível construir mapas de saliência em escalas mais baixas da imagem para economizar tempo de processamento. Além disso, a base foi produzida apenas com 8 indivíduos por mapa, gerando poucos pontos por mapa, o que provoca uma análise estatística pobre para a análise desejada nesse trabalho, além de não ser a melhor quantidade de pessoas para a publicação da base, como discutido pelo trabalho posterior do MIT na base MIT300 [33], afirmando que um número mais coerente seria de 39 pessoas por mapa.

Discutidas as principais bases de mapas de densidade e fixacão da literatura, a próxima seção faz uma revisão das principais abordagens das técnicas de extração de características locais.

2.4 Extração de Características

Em visão computacional o termo "característica" se refere a uma parte de uma imagem com alguma propriedade especial, por exemplo, linhas, cantos, bordas, áreas de alto constraste de cor e regiões de textura [34]. Essas características são amplamente exploradas por diversos pesquisadores da visão computacional, pois as aplicações referentes a reconhecimento de objetos, odometria visual, mapeamento, detecção de movimentos, correspondência entre imagens e reconstrução de cenas 3D dependem da presença de características estáveis e representativas.

No decorrer dos últimos 50 anos, surgiram diversas abordagens referentes à extração de características em uma imagem. As abordagens mais utilizadas se referem aos algoritmos detectores e descritores de características locais. Uma característica local é um padrão de imagem que difere de sua vizinhança imediata. Geralmente está associado a uma alteração de uma propriedade da imagem ou várias propriedades simultaneamente. As características locais geralmente podem ser representadas por pontos, bordas ou pequenos *patches* da imagem.

Nesse sentido, os detectores de características locais têm como função buscar e

encontrar a localização de pontos, bordas ou até mesmo pequenas regiões da imagem que representem bem as características buscadas. Por conta disso, esses algoritmos procuram ser robustos a diversas transformações em imagens, como mudanças de pontos de vista, escala, rotação, iluminação e borramento [35]. Essa preocupação é necessária para aplicações práticas, pois o algoritmo precisa suportar essas variações detectando pontos estáveis na imagem.

Encontrados os elementos que representam essas características, sejam eles pontos, bordas, ou *patches*, é tarefa dos algoritmos de descrição dar uma identidade aos elementos encontrados pelo detector. É desejável que essa identidade seja única para que os elementos possam ser facilmente distinguíveis automaticamente, sendo possível assim encontrar os mesmos elementos na mesma cena mesmo ela passando por diversas transformações de escala, rotação, perspectiva, luminosidade, compressão, borramento e translação, por exemplo.

Para um algoritmo ser um bom de detector e descritor de características locais ele deve atender alguns requisitos. Segundo [36], os principais requisitos são:

- ***Repetibilidade:*** Dadas duas imagens do mesmo objeto ou cena, tiradas sob diferentes condições, uma alta porcentagem de características devem ser detectadas igualmente em ambas as imagens. Essa característica está ligada à robustez e à invariância que as técnicas devem possuir com relação às transformações que podem ocorrer em uma imagem.

- ***Distinção:*** As características detectadas devem mostrar muita variação, de modo que possam ser diferenciadas e combinadas.

- ***Localidade:*** As características devem ser locais, de modo a reduzir a probabilidade de oclusão e permitir aproximações simples do modelo das deformações geométricas e fotométricas entre duas imagens tiradas sob diferentes condições de visualização.

- ***Quantidade:*** O número de características detectados deve ser suficientemente grande, mesmo em pequenos objetos. No entanto, o número ideal depende da aplicação. Idealmente, a densidade de características deve refletir o conteúdo da informação da imagem para fornecer uma representação compacta da imagem.

- ***Acurácia:*** As caraterísticas detectadas devem ser localizadas com precisão, tanto na localização da imagem quanto no que diz respeito à escala e possivelmente à forma.

- ***Eficiência:*** Sempre que possível a detecção de características em uma nova imagem deve permitir aplicações de tempo crítico.

Dados os principais requisitos desejados para as técnicas de extração de características, as Seções 2.4.1 e 2.4.2 descrevem de modo mais detalhado os principais detectores e descritores encontrados na literatura.

2.4.1 Detectores Locais de Características

As pesquisas com relação a detectores de características ocorrem desde 1954, quando foi observado pela primeira vez por Attneave [37] que a informação sobre a forma está concentrada em pontos de cantos e junções. A partir dessas descobertas já foi percebido que no campo de processamento de imagens as linhas e cantos têm um papel importante nos algoritmos de detecção. A partir de então, essa área se desenvolveu em várias direções.

Vale ressaltar que os experimentos feitos por Attneave em 1954 foram realizados ainda de forma bastante arcaica, na época não haviam computadores modernos e de fácil acesso, muito menos equipamentos de rastreamento ocular para se ter uma análise mais profunda do comportamento da percepção visual humana. Apesar disso, ele conseguiu ter vestígios muito bons de alguns elementos que contribuem para o reconhecimento de objetos mesmo que de forma voluntária através do sistema visual humano, mas não de todos os elementos que podem existir para a realização dessa tarefa. Na verdade até a atualidade ainda não existe um número fechado de características da imagem que pode ser útil para essa tarefa.

Com as descobertas de Attneave uma série de direções de pesquisa foram criadas na área de extração de características. Primeiro, muitos autores estudaram a curvatura de contornos para encontrar cantos. Outros analisaram diretamente as intensidades da imagem, buscando regiões de alta variância. Foram criados também métodos focados na exploração da informação das cores. Mais recentemene foram criados métodos de detecção com invariância, métodos baseados em técnicas de aprendizado de máquina e métodos com alguma inspiração no sistema visual humano. Diante de tantas abordagens existentes, de forma qualitativa, as técnicas podem ser divididas em três grupos: detectores de cantos, detectores de bordas e detectores de *blobs* (região).

Essas abordagens utilizam de alguma forma a informação do gradiente local na detecção, já que uma borda é uma divisa entre duas regiões pertecentes a uma imagem, ocorrendo justamente pelo alto gradiente entre duas regiões. Os cantos diferenciam-se das bordas apenas por terem direções diferentes de gradiente, formando assim uma curvatura mais acentuada, já os *blobs* detectam uma região discriminante da imagem por altos gradientes variando em todas as direções.

Como existem várias abordagens de detecção de características locais, apenas algumas das técnicas mais relevantes para a área são melhor detalhadas no decor-

rer desta seção. Para finalizar, é feita uma discussão mostrando tecnicas que se desenvolveram posteriormente, baseadas nas tecnicas mais importantes, e onde se encontra o estado da arte da área.

Harris Corner Detector

O Harris Corner Detector e suas versões aperfeiçoadas, o Harris-Laplace e o Harris-Affine são a família de detectores mais representativos do grupo de detecção de cantos. Esses cantos não correspondem necessariamente na imagem 2D a projeções de cantos em 3D. Na verdade, esses cantos podem também ser vários tipos de junções, superfícies com elevadas curvaturas e superfícies altamente texturizadas.

O Harris Corner Detector basea-se na matriz de autocorrelação. Essa matriz descreve a distribuição de gradiente em uma vizinhança local de um ponto. Os autovalores dessa matriz representam as principais mudanças de sinal em duas direções ortogonais em uma vizinhança ao redor de um ponto. Com base nessa propriedade, os cantos podem ser localizados como locais na imagem para os quais o gradiente varia significativamente, ou em outras palavras, para os quais os autovalores da matriz são grandes.

Foi descoberto que autovalores da matriz de autocorrelação podem ajudar a determinar quais pontos são bons cantos. Uma pontuação R é então calculada para cada matriz construída em torno de um ponto:

$$R = det(M) - k \times traço(M) \tag{2.1}$$

em que o $det(M) = \lambda_1\lambda_2$ e o $traço(M) = \lambda_1 + \lambda_2$. Sendo que λ_1 e λ_2 são os dois maiores autovalores da matriz M e k é um valor constante tipicamente usado como 0.04.

Desse modo, as janelas de pontos cujo valor R é maior que um determinado limiar são considerados bons pontos e são considerados pelo sistema de detecção.

Os pontos encontrados por esse detector têm uma boa invariância rotacional e de iluminação, mas não são invariantes à escala. Para suprir essa necessidade, o Harris-Laplace foi criado. Ele se trata basicamente de uma modificação na matriz de autocorrelação para refletir a busca de pontos no espaço de escala utilizando kernels gaussianos. Já o detector Harris-Affine faz modicações na matriz de autocorrelação do Harris-Laplace para tornar a detecção invariante à transformações afins.

Por conta da estabilidade dessa família de detectores e da precisão na localização dos pontos eles são bastante utilizados em tarefas de calibração de câmeras e reconstrução 3D.

SUSAN

O detector de canto SUSAN foi introduzido por Smith e Brady [38] e conta com uma técnica diferente. Em vez de avaliar gradientes locais, que podem ser sensíveis ao ruído e computacionalmente mais caros, uma abordagem morfológica é usada.

Diferentemente da família de detectores Harris, o SUSAN realiza a detecção de cantos com uma técnica mais simples. Para cada *pixel* da imagem é considerada uma vizinhança circular de raio fixo. O *pixel* central é chamado de núcleo e seu valor de intensidade é usado como referência. Assim, todos os *pixels* dentro dessa vizinhança são divididos em duas categorias, a categoria de valores similares ao núcleo e a categoria de valores diferentes. A proporção de pontos existentes nas duas categorias reflete o tipo de área da imagem que está sendo analisado. Em regiões muito homogêneas a porcentagem de pontos semelhante ao núcleo cobre praticamente toda a área circular, enquanto que em regiões de borda essa proporção cai para 50% e em cantos para 25%. Seguindo esses limiares, os cantos são detectados como locais na imagem onde a porcentagem do número de *pixels* está abaixo de um limite predefinido.

Enquanto a família de detectores Harris é mais estável e uma ferramenta muito conveniente quando se deseja obter muitos pontos por imagem, o detector SUSAN é mais eficiente, mas também mais sensível ao ruído.

FAST

O FAST (*Features from Accelerated Segment Test*) foi proposto por Rosten e Drummond [39] e é um aperfeiçoamento do detector de cantos baseado no SUSAN.

Assim como o SUSAN, o FAST utiliza o centro de uma área circular para determinar a intensidade de todos os seus *pixels* vizinhos. No entanto, o FAST não analisa toda a área circular, somente *pixels* que estão em um raio r de distância, gerando uma máscara de segmento circular que acelera o teste para identificação de pontos, e é conhecida como *Accelerated Segment Test* (AST). Essa máscara é fundamental para o FAST, já que ele foi criado justamente para acelerar a detecção de pontos de interesse para facilitar o uso de detectores em aplicações de tempo real. A máscara desenvolvida nessa abordagem é ilustrada na Figura 2.11.

Como ilustrado na Figura 2.11 o *pixel* central P é comparado com todos os 16 *pixels* do segmento pertencente a um raio r. Se 9 *pixels* contíguos (pontilhado azul claro) não estiverem entre o intervalo $P+t$ e $P-t$, onde P é a intensidade do pixel central e t um limiar, então P é candidato a um ponto de interesse. Para acelerar a busca primeiro são testados os *pixels* 1 e 9 (círculos verdes); se suas intensidades não estiverem entre o intervalo $P+t$ e $P-t$, os *pixels* 5 e 13 (círculos azuis) são verificados também. Se três dos quatro passarem no teste, os demais são verificados

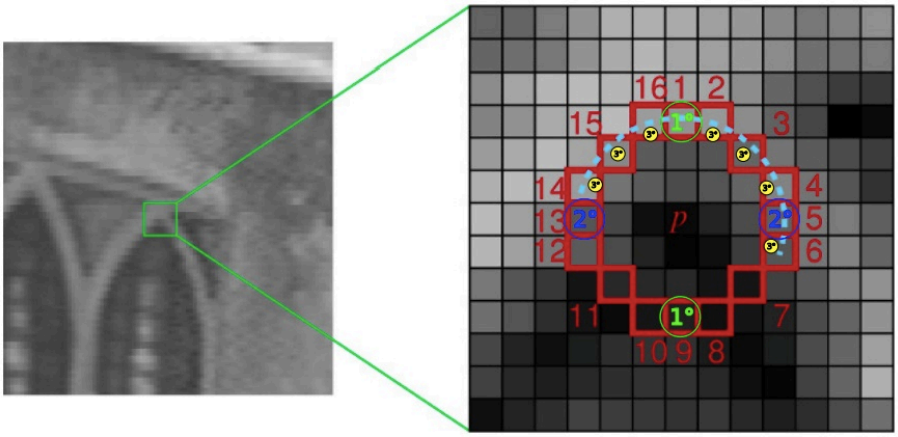


Figura 2.11: Máscara AST utilizada no FAST.

para identificar um segmento de 9 *pixels* que passe pelo critério. De forma geral, um ponto de interesse é determinado dado um conjunto de 9 *pixels* conectados pertencentes ao segmento de 16 *pixels*, com intensidades maiores ou menores que um limiar, determinado pela intensidade do pixel central. O limiar determina a quantidade e qualidade (magnitude do gradiente) dos pontos encontrados.

O uso dessa máscara traz quatro problemas:

- Para um conjunto contíguo menor que 12 *pixels*, muitos pontos de interesse são detectados.
- A escolha da ordem dos pixels para a primeira comparação assume que a distribuição da aparência do ponto de interesse é conhecida, influenciando diretamente na rapidez do algoritmo.
- Os critérios de teste que agilizam na validação de um possível ponto de interesse são descartados.
- Muitos pontos são detectados adjacentes uns dos outros.

Os primeiros três problemas foram resolvidos com uma técnica de aprendizado de máquina utilizando o algoritmo de árvore de decisão ID3. Nessa técnica são selecionados conjuntos de imagens para treinamento. Para cada imagem o algoritmo FAST é executado para encontrar os candidatos a pontos de interesse junto com os vetores de 16 *pixels* que delimitam a máscara de cada ponto. Para cada ponto candidato encontrado é armazenado o seu respectivo vetor de vizinhos de 16 elementos. Esse processo é repetido para todos os *pixels* para todo o conjunto de imagens de treinamento. A partir de então, é criado um vetor P com todos os dados de treinamento. Cada ponto do vetor de 16 elementos pertecentes à vizinhança de um ponto candidato pode pertecer a três estados diferentes. O ponto pode ser mais escuro, mais claro ou considerado similar ao candidato. Matematicamente, dado por:

$$S_{p\to x} = \begin{cases} e, & I_{p\to x} \leq I_p - t & (escuro) \\ s, & I_p - t < I_{p\to x} < I_p + t & (similar) \\ c, & I_p + t \leq I_{p\to x} & (claro) \end{cases} \tag{2.2}$$

onde $S_{p\to x}$ é o estado, $I_{p\to x}$ é a intensidade do pixel x, e t é um limiar.

A partir de então, com posse de todos os vetores que são pontos de interesse ou não, a árvore é treinada utilizando o princípio da minimização da entropia, e retorna se o *pixel* central é um ponto de interesse ou não, consultando o vetor de 16 *pixels* o menor número de vezes possível. Esta árvore ternária é treinada para cada novo ambiente.

O quarto problema é resolvido com uma supressão de não máximos. Na supressão de não máximos é calculada uma função de pontuação V para cada ponto detectado. Essa pontuação é a soma da diferença absoluta de intensidade entre o ponto detectado e os 16 *pixels* adjacentes a ele. Após isso, dois pontos-chaves adjacentes são comparados e é descartado o ponto com menor valor de V.

AGAST

Como o FAST apresenta a desvantagem da árvore precisar ser treinada para cada novo ambiente, surgiu o detector AGAST (*Adaptive and Generic Accelerated Segment Test*) em 2010 para corrigir esse problema. Nessa nova abordagem é implementada uma árvore binária genérica que não necessita ser retreinanda para cada novo ambiente [40].

O AGAST é baseado no mesmo critério AST utilizado no FAST, mas usa uma árvore de decisão diferente. O AGAST é treinado com base em um conjunto de dados que inclui todas as combinações possíveis de 16 *pixels* no círculo. Isso garante que a árvore de decisão funcione em qualquer ambiente. Além disso, a AGAST introduz um algoritmo dinâmico de troca de árvore, que altera automaticamente as árvores de decisão. Uma árvore é treinada em áreas homogêneas e outra é treinada em áreas heterogêneas. Desta forma, o desempenho do AGAST aumenta para cenas aleatórias. Combinando essas duas melhorias, o AGAST trabalha em qualquer ambiente sem necessidade de retreinamento.

SIFT

O SIFT (*Scale Invariant Feature Transform*) foi criado por Lowe em 1999 e é um método que detecta pontos de interesse extremamente distintos, têm alta repetibilidade e são invariantes a rotação, escala, pontos de vista e iluminação. O SIFT associa escala e orientação por meio de dois estágios, detecção e descricão,

sendo que descrição será abordada na seção 2.4.2 de descritores.

O estágio de detecção do SIFT pode ser dividido por três estágios: detecção de extremos no espaço de escala, localização de pontos de interesse e orientação de pontos de interesse.

A primeira etapa o método utiliza a Diferença de Gaussianas (DoG) para identificar possíveis pontos de interesse. O espaço de escala é definido como uma função $L(x,y,\sigma)$ produzida pela convolução de um filtro gaussiano $G(x,y,\sigma)$ em uma imagem $I(x,y)$, ou seja:

$$\begin{aligned} L(x,y,\sigma) &= G(x,y,\sigma) * I(x,y) \\ G(x,y,\sigma) &= \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \end{aligned} \tag{2.3}$$

Para que o cálculo seja mais simples e eficiente computacionalmente, é utilizado um fator de escala γ para variar a distribuição do filtro Gaussiano entre as imagens, de modo que a diferença entre estas é uma aproximação ao Laplaciano da Gaussiana (LoG) feito para encontrar os extremos $D(x,y,\sigma)$ conforme a equação:

$$\begin{aligned} D(x,y,\sigma) &= (G(x,y,\gamma\sigma) - G(x,y,\sigma)) * I(x,y) \\ D(x,y,\sigma) &= L(x,y,\gamma\sigma) - L(x,y,\sigma) \end{aligned} \tag{2.4}$$

A Figura 2.12 ilustra os passos para detecção de pontos de interesse. Primeiro é construída a pirâmide da imagem original por oitavas. Após, a imagem original de cada oitava é borrada por um filtro Gaussiano multiplicado por um fator γ para gerar as escalas (Figura 2.12 (a)). Em seguida é calculada a diferença entre as escalas, para cada oitava da pirâmide, gerando uma imagem aproximada ao filtro LoG (Figura 2.12 (b)). Por fim, cada pixel da imagem gerada pela diferença (DoG) (Figura 2.12 (c) marcada com x) é comparado com seus oito vizinhos e com os 9 pixels da escala superior e inferior, para saber se este é um extremo máximo ou mínimo, tornando-se então um possível candidato a ponto de interesse.

A estes extremos é aplicada uma aproximação quadrática local para identificar a posição do pico ou vale. A aproximação quadrática é feita tomando uma expansão de Taylor em torno do ponto atual. Isso fornece uma aproximação de posição que tem resolução de *subpixel* e uma estimativa de escala bastante precisa.

Na segunda etapa são feitos alguns tratamentos para eliminar pontos de interesse cujo gradiente dos pixels vizinhos são baixos ou que encontram-se ao longo de bordas.

A Figura 2.13 em (a) evidencia todos os pontos de máximo e mínimo encontrados na primeira etapa de detecção. A Figura 2.13 em (b) mostra os pontos que permaneceram na imagem após a etapa de eliminação de pontos de baixo gradiente.

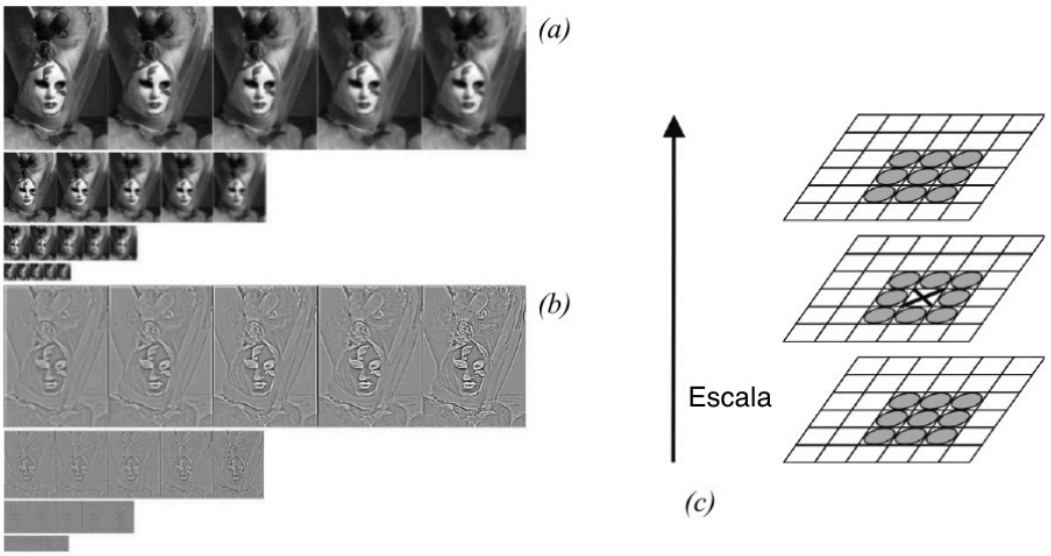


Figura 2.12: Detecção de extremos no espaço de escala do SIFT. Em (a), as imagens na vertical representam a pirâmide feita por oitavas, e na horizontal os borramentos Gaussianos de diferentes tamanhos de filtro; em (b) é apresentada a diferença entre imagens Gaussianas de (a) em todas as oitavas; e em (c) é comparado o pixel central com seus oito vizinhos, incluindo as escalas, para identificar extremos máximos e mínimos.

Por fim, a Figura 2.13 em (c) mostra os pontos que permaneceram na imagem após a remoção dos pontos de borda. Estes procedimentos retornam um grupo de pontos de interesse que são localizados com acurácia de *subpixel* e que foram detectados em uma determinada escala.

A terceira etapa finaliza o processo de detecção de pontos de interesse atribuindo a cada ponto uma orientação. Esta orientação é calculada analisando a intensidade do gradiente (magnitude e orientação) de cada *pixel* vizinho. Um histograma das orientações é construído fazendo com que cada valor do gradiente de cada pixel contribua para formação do valor final. Um pico no histograma torna-se a orientação

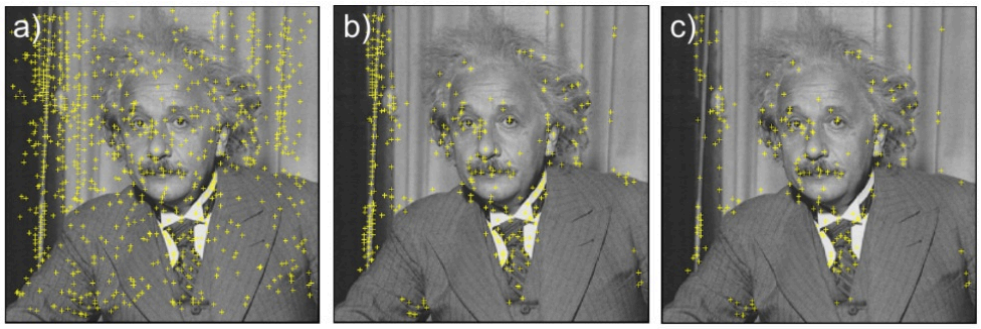


Figura 2.13: Pontos de interesse localizados utilizando o SIFT. Em a), todas os pixels máximos e mínimos são representados por pontos amarelos; em b), são eliminados os pontos que estão em regiões de baixo gradiente; e em c) os pontos instáveis em bordas são eliminados conforme valores de resposta da Matriz Hessiana.

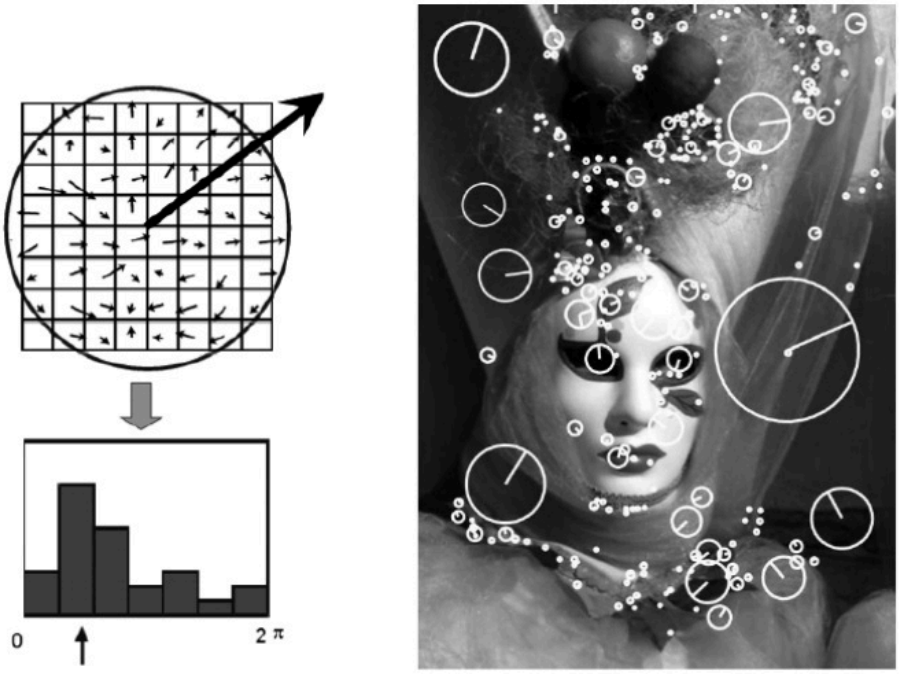


Figura 2.14: Orientação do ponto de interesse encontrado pelo SIFT. Em (a), a intensidade do gradiente (magnitude e orientação) de todos os pixels vizinhos ao pixel central são calculadas e após, representadas em forma de histograma conforme orientação. Um pico detectado no histograma torna-se a orientação dominante no ponto. Ainda, o valor atribuído a orientação do ponto é ponderado pela escala onde este foi detectado; em (b) é apresentado um exemplo de pontos detectados em uma imagem, conforme sua escala e orientação.

dominante no ponto e, se mais de um pico surgir, um ponto de interesse adicional é criado para cada pico contendo localização e escala do ponto original mensurado. O valor atribuído à orientação do ponto é ponderado pela escala onde foi encontrado (Figura 2.14 (a)). Assim, todas as propriedades dos pontos de interesse serão mensuradas com relação à sua orientação, os tornando então invariantes à escala e rotação (Figura 2.14 (b)) [35].

As equações que definem a magnitude e orientação são:

$$\begin{aligned} M(x,y) &= \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \\ \theta(x,y) &= tan^{-1}(\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}) \end{aligned} \tag{2.5}$$

onde M é a magnitude do pixel e θ é a orientação.

A terceira etapa contribui para a redução da sensibilidade em relação a mudanças do ponto de vista da câmera e alterações não lineares no brilho da imagem (variações lineares são removidas pelas operações de gradiente), analisando regiões na locali-

dade do ponto de interesse. Em Lowe [35] esta técnica de detecção é abordada com muito mais detalhes além de descrever fatores que incrementam seu desempenho.

SURF

O SURF (*Speeded Up Robust Features*) foi criado por Bay et. al e se trata de um método baseado nos mesmos princípios e passos utilizados pelo SIFT, porém com esquemas diferentes que agilizam o processo sem perder a robustez [41].

A detecção de pontos de interesse do SURF usa a Matriz Hessiana assim como o Harris, para extrair os autovalores e caracterizar o ponto pelo determinante. A imagem integral proposta em Viola e Jones conjuntamente com as janelas de filtro proposto em Simard et al. em 1999 [42], reduzem drasticamente o tempo de processamento por necessitarem de menos cálculos para determinar o gradiente de uma região. Estes esquemas são a base do SURF e serão descritos a seguir.

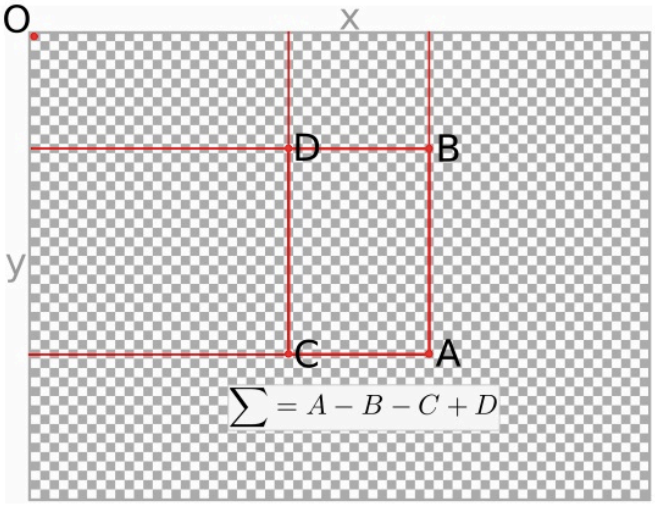


Figura 2.15: Técnica de imagem integral utilizada pelo SURF. A letra (O) representa o início das coordenadas (x) e (y) de uma imagem onde os quadrados brancos e cinzas são posições de *pixels*; em (D) é armazenada a soma acumulativa da intensidade de todos os pixels pertencentes ao retângulo cuja diagonal vai do ponto O ao ponto D, assim como qualquer coordenada da imagem (A, B ou C); então para se calcular o somatório das intensidades de uma região quadrada qualquer, basta acessar os valores de quatro coordenadas (A, B, C e D) da imagem integral e executar três operações (A - B - C + D).

O conceito de imagem integral é armazenar o somatório cumulativo de todos os pixels de uma área retangular (Figura 2.15, pontos A, B, C e D), onde o *pixel* da extrema direita inferior (ponto A) acaba por ter o resultado do somatório de todos os *pixels* pertencentes ao maior retângulo da imagem da Figura 2.15, ou seja, o retângulo cuja diagonal vai do ponto O ao ponto A. Logo, a imagem integral na coordenada $x' = (x, y)$ representa a soma de todos os *pixels* da imagem original na região retangular de início em O e fim em x' conforme a Figura 2.15.

$$I_{\sum(x')} = \sum_{i=0}^{i \leq x} \sum_{j=0}^{j \leq y} I(i,j) \tag{2.6}$$

Depois que a imagem integral estiver calculada, são necessárias somente três adições e quatro acessos a memória para extrair a soma das intensidades dos pixels da área retangular. O tempo de cálculo independe do tamanho do retângulo, o que é importante para o SURF, devido ao tamanho das janelas de filtro utilizados para aproximar a Matriz Hessiana.

Para calcular a Matriz Hessiana é necessário calcular derivadas parciais de segunda ordem nas direções x (horizontal), y (vertical) e xy (diagonal) por meio de janelas de filtro Gaussiano (Figura 2.16 centro esquerda). O SURF aproxima estas derivadas por filtros de Haar (Figura 2.16 centro direita), onde a região cinza é igual a zero.

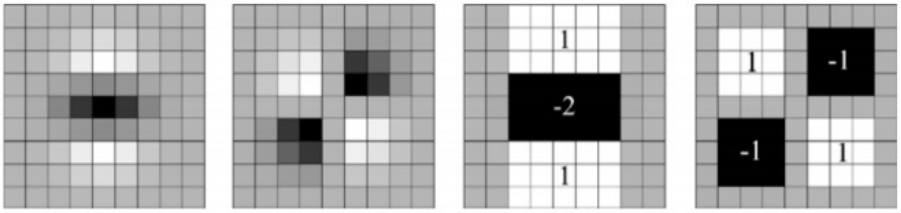


Figura 2.16: Filtros Gaussianos (centro esquerda) e Filtros Haar (centro direita). Os *kernels* centro esquerda representam filtros com distribuições Gaussianas de pesos; os filtros centro direita são conhecidos como Haar e são aproximações dos filtros Gaussianos, estes em conjunto com a imagem integral agilizam o processo de derivação.

Ao convolver na imagem uma janela de filtro 9×9, o resultado gerado representa uma distribuição Gaussiana aproximada com desvio padrão $\sigma = 1,2$ que é a menor escala utilizada para encontrar o ponto característico. Como cada escala aproximada tem um valor de σ diferente, um peso w constante é utilizado para simplificar a equação e manter a energia entre a janela Gaussiana e a janela Gaussiana aproximada. Logo, denotando as derivadas parciais na horizontal como D_{xx}, na vertical como D_{yy} e diagonal como D_{xy}, pode-se calcular o determinante aproximado *Det(Haprox)* para qualquer tamanho de janela de filtro conforme Equação 2.7.

$$Det(Haprox) = D_{xx}D_{xy} - (wD_{xy})^2 \tag{2.7}$$

O peso w é sensível à escala mas pode ser mantido com o valor de *0,9* sem ter impacto no resultado, conforme Equação 2.8.

$$w = \frac{|L_{xy}(1,2)|_F |D_{yy}(9)|_F}{|L_{yy}(1,2)|_F |D_{xy}(9)|_F} = 0.912... \simeq 0.9 \tag{2.8}$$

onde $||_F$ é a norma de Frobenius, L a derivada de segunda ordem Gaussiana e D a derivada aproximada pela janela de filtro do SURF.

Para que o SURF seja invariante à escala, é necessária a criação de oitavas como SIFT para encontrar o máximo local. O processo inicia convolvendo uma janela de filtro 9×9 para a menor escala, depois a janela é aumentada para 15×15, 21×21 e 27×27, criando assim quatro escalas para a primeira oitava. Na segunda oitava os filtros aumentam (15, 27, 39, 51). Na terceira para (27, 51, 75, 99) e quarta (51, 99, 147, 195). Diferentemente do SIFT que diminui pela metade a imagem original em cada oitava, o SURF mantêm as dimensões da imagem original e aumenta para o dobro o tamanho dos filtros. O processo de busca do máximo local é igual ao do SIFT. Maiores detalhes a respeito do processo de aumento do tamanho dos filtros podem ser encontrado em [41].

Estado da Arte dos Detectores Locais

Diversas idéias surgiram principalmente baseadas no detectores de cantos e no SIFT. Basicamente as abordagens seguiram dois caminhos: aquelas que são baseadas em características identificadas manualmente pelo projetista e aquelas indentificadas através de técnicas de aprendizado de máquina, sendo essas as mais recentes e as que vêm sendo mais investigadas atualmente.

No grupo de técnicas com características identificadas manualmente surgiram abordagens posteriores aos algoritmos mais clássicos, como o detector MSER [43], o CenSurE [44], o detector de máximos de intensidade de VanGool [36], o ASIFT [45], o detector de máximos de entropia proposto por Kadir e Brady [46] [47] e muitos outros detectores que têm como propósito encontrar regiões de borda, cantos ou *blobs*. Uma visão mais detalhada de outros detectores clássicos pode ser encontrada em [36].

No grupo de técnicas mais recentes o aprendizado de máquina é utilizado para definir quais serão os pontos encontrados pelo detector. Atualmente essa tem sido uma questão para novas pesquisas e novas discussões, pois os detectores tradicionais conseguem características de invariância por focarem sua detecção em elementos da imagem como cantos e *blobs*, que sofrem poucas alterações em sua estrutura com as modificações da cena. Além disso, não há uma lista completa de quais estruturas em uma imagem apresentam características de invariância. Assim, o desafio das abordagens de aprendizado é aprender e determinar quais elementos na cena têm propriedades de invariância e determinar onde estão os alvos de detecção.

Nessa área o principal representante é o trabalho proposto por Karel Lenc et. al. [48]. Nele, é introduzida uma formulação de aprendizado para detectores baseados em duas idéias: (i) definir uma função objetivo em termos de uma restrição de covariância que é independente do tipo de elemento analisado na imagem e (ii) formular a detecção como um problema de regressão, o que permite usar poderosos regressores como redes profundas para esta tarefa.

Seguindo essa abordagem, em [49] foram introduzidos a essa técnica conceitos de "patch padrão" e "características canônicas" nas etapas de treino do detector, contribuindo para a construção de um detector mais robusto e com uma repetibilidade maior. Um outro trabalho bastante relevante nessa área é a abordagem chamada LIFT (*Learned Invariant Feature Transform*) [50] que implementa uma arquitetura chamada de *DeepNetwork* que realiza a tarefa de detecção, estimação de orientação e descrição de características de maneira unificada.

2.4.2 Descritores Locais de Características

Na Seção 2.4.1 foram apresentados como os pontos de interesse são detectados. No entanto, para que esses pontos sejam comparados, é necessário que exista uma forma de descrevê-los. A abordagem usual para esse problema é utilizar vetores de características para construir os descritores.

Na literatura são encontrados diferentes tipos de descritores, e não há um consenso a respeito de uma nomenclatura de divisão das abordagens existentes. A divisão mais encontrada na literatura foi proposta por Mikolajczyk [51]. Nessa abordagem os descritores são divididos entre: baseados em distribuição de *pixels*, abordagens diferenciais, abordagens baseadas em frequência espacial e um grupo de outras técnicas onde ficam os métodos baseados em invariantes de momento, utilizados para descrever objetos. No entanto, essa divisão é mais antiga e não consegue agrupar as novas técnicas que surgiram nos últimos 10 anos na área de descrição de características. Por isso, nesse trabalho os descritores serão divididos entre dois grupos, que serão melhor apresentados no decorrer dessa seção, sendo eles: descritores de ponto flutuante e descritores binários.

Descritores de Ponto Flutuante

O descritor mais simples é o vetor de *pixels* em torno do ponto de interesse da imagem. A partir daí, a correlação cruzada é utilizada para calcular uma pontuação de similaridade entre os vetores de pontos. No entanto, a alta dimensionalidade de tal descrição resulta em uma alta complexidade computacional para o reconhecimento. Por isso algumas técnicas costumam aplicar o PCA para reduzir o tamanho do descritor. Para reduzir a dimensionalidade, outra técnica utilizada é a constru-

ção de um histograma com a distribuição de intensidades dos *pixels* na vizinhança do ponto de interesse. Essa representação é mais eficiente, mas pouco robusta a algumas transformações, como a de iluminação, por exemplo.

Com o aumento da demanda por aplicações cada vez mais robustas, se tornou necessário o desenvolvimento de descritores mais robustos às transformações, o que não é possível alcançar apenas com a informação bruta da intensidade. Foram então desenvolvidos descritores que utilizam relações mais complexas utilizando os valores dos *pixels* vizinhos de um ponto de interesse. O descritor utilizado pelo SIFT [35] surgiu em resposta à observação dessas deficiências, construindo um descritor invariante à rotação, iluminação e pontos de vista 3D.

Para descrever o ponto de interesse, o SIFT calcula primeiramente a magnitude e orientação do gradiente dos *pixels* que são amostrados ao redor da localização do ponto-chave. Este procedimento é ilustrado na Figura 2.17 em (B), onde os gradientes são representados pelas pequenas setas pretas e a região amostrada é uma área de 8×8 *pixels* dividida em 4 regiões de dimensão 2×2, em que cada região possui 4×4 *pixels*. Os valores encontrados nessa fase são normalizados com relação à orientação obtida na fase de detecção, obtendo assim a invariância a rotação.

Uma função Gaussiana é utilizada para dar peso à magnitude do gradiente em cada ponto na vizinhança do ponto-chave, com uma janela de suavização Gaussiana de escala σ igual à metade da largura da janela do descritor. Essa Gaussiana evita mudanças súbitas do descritor e reduz a ênfase nos gradientes longe do ponto no centro do descritor, evitando maior propagação de erros.

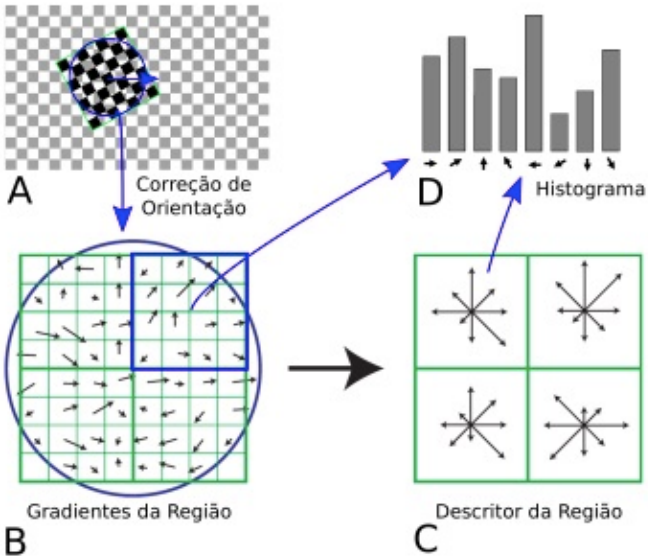


Figura 2.17: Etapas de construção do descritor SIFT.

Efetuada a suavização dos gradientes, o descritor consiste de um vetor contendo os valores de 4 histogramas empilhados um após o outro. Um histograma (Figura

2.17), é criado para cada região da matriz 2×2 inicialmente amostrada em torno do ponto-chave. Cada região possui 16 elementos, cuja orientação e magnitude suavizadas pelas Gaussianas são utilizadas para construir um histograma em que o eixo x possui até 8 orientações e o eixo y é representado pela magnitude do gradiente, como ilustrado na Figura 2.17. Construído o histograma para as 4 regiões da janela 2×2, o vetor descritor resultante passa a ser produto da junção de 4 histogramas de 8 elementos cada, resultando em um vetor de 32 características. No entanto, o vetor de 32 características ainda é pouco descritível, por isso, Lowe adota uma janela de 4×4 em sua pesquisa, resultando em um descritor de $4 \times 4 \times 16 = 128$ elementos.

Para que o descritor tenha invariância a iluminação ele é normalizado, tornando-se invariante a mudanças homogêneas. Já com relação a mudanças não-lineares, causadas pela saturação de câmeras ou por efeito de iluminação de superfícies tridimensionais em diferentes orientações, isso pode provocar elevada influência sobre as magnitudes dos gradientes, mas pouca influência sobre a orientação. Esse efeito é reduzido impondo um valor máximo às magnitudes. Após a normalização, todos os valores acima de um determinado limiar são ajustados para este limiar, isso faz com que direções com magnitudes muito grandes não dominem a representação do descritor. Maiores detalhes a respeito dos limiares e de etapas de construção do descritor SIFT podem ser encontrados em [35].

O sucesso trazido pelo SIFT fez dele um dos algoritmos de extração de características mais citado e utilizado nos últimos anos. Mikolajczyk e Schmid [51] avaliaram uma variedade de descritores e identificaram o descritor SIFT como sendo o mais resistente a deformações comuns de imagem, mas com um custo computacional elevado. A fim, de promover melhorias no custo computacional, alguns anos depois foi criado o PCA-SIFT [52].

O PCA-SIFT faz uso da técnica de Ánalise de Componentes Principais (PCA) para reduzir o espaço de armazenamento de um vetor de características e o tempo de cálculo de distâncias para encontrar correspondências. Ele utiliza o mesmo princípio de cálculo de gradiente. Entretanto, utiliza uma malha de tamanho 39×39 em vez de 4×4, formando um vetor de 3042 dimensões. Esse vetor é reduzido a 36 componentes com ajuda do PCA. Essa técnica se mostrou robusta, a princípio, mas perde em precisão para o SIFT em alguns casos específicos [52].

Uma outra alternativa para reduzir o tempo de processamento do SIFT foi a criação do SURF, que desde as etapas de detecção reduz a quantidade de operações necessárias. Na etapa de descrição também não é diferente. Nela, o SURF utiliza uma transformada de Haar nas direções x e y (Figura 2.18 (A)) para encontrar intensidades em todos os *pixels* vizinhos ao ponto, a uma distância de raio proporcional à escala em que o ponto foi encontrado. Para calcular a resposta do filtro de Haar em qualquer das direções x e y, somente seis operações são necessárias em qualquer

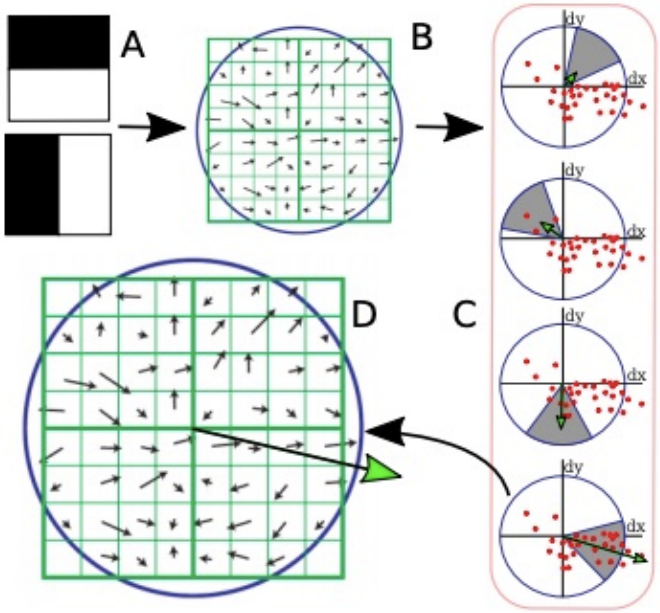


Figura 2.18: Orientação do ponto dominante - SURF.

escala devido à imagem integral. Um peso Guassiano é utilizado para incrementar a região circular ao ponto de interesse e ponderar as respostas obtidas pela filtragem (Figura 2.18 (B)). Esse resultado é representado por pontos no espaço em que as intensidades horizontais aparecem ao longo do eixo x, e verticais ao longo do eixo y (Figura 2.18 (C)). A orientação dominante é calculada somando todos os pontos que encontram-se em uma janela de tamanho $\frac{\pi}{3}$. A orientação dominante é ilustrada na Figura 2.18 em (D).

Uma vez encontrada a orientação dominante, uma região quadrada é construída ao redor do ponto de interesse e orientada conforme o mesmo, em que o tamanho da janela é 20 vezes a escala utilizada (Figura 2.19 (A)). A janela é dividida regularmente em sub-regiões quadradas de 4×4 *pixels* (Figura 2.19 (B)). Para cada sub-região 5×5 a transformada de Haar é calculada 25 vezes e são coletadas as orientações relacionadas às transformadas em x com o nome dx e em y com o nome dy (Figura 2.19 (C) e (D)). Para aumentar a robustez são atribuídos pesos de uma Guassiana centrada no ponto de interesse. O somatório de dx e dy é feito em cada sub-região e é armazenado no vetor de características do ponto de interesse. Desta forma, para cada sub-região existe um vetor que representa suas intensidades sendo ele composto por: $\sum dx, \sum dy, \sum |dx|, \sum |dy|$ (Figura 2.19 (E)), e concatenando todas as 4×4 sub-regiões é obtido o vetor descritor com 64 características.

Em [51] foi proposto um novo descritor baseado nas ideias do SIFT e do PCA-SIFT, o *Gradient Location and Orientation Histogram* (GLOH). Ele é um descritor que calcula histogramas de gradientes em uma grade circular (Figura 2.20) com 17 sub-regiões e 16 bins de orientação por sub-região, formando um vetor com 272 di-

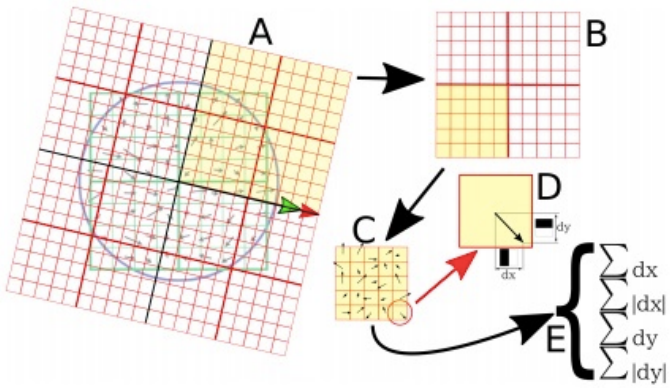


Figura 2.19: Etapas de construção do descritor SURF.

mensões. Por fim, o vetor de características é reduzido para 128 dimensões usando o PCA. Posteriormente, Dalal e Triggs [53] criaram o descritor HOG que combina propriedades tanto do SIFT como do GLOH. Ele também é representado por um histograma de localização de gradientes e orientações. A principal diferença entre o HOG e o SIFT é que o HOG é calculado em uma grade densa de células uniformemente espaçadas, gerado sobreposições do contraste local.

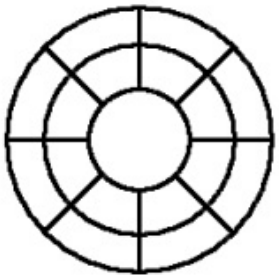


Figura 2.20: Grade circular usada para calcular os histogramas de gradientes do GLOH.

Foram criadas técnicas inspiradas no SIFT também diretamente em imagens coloridas. Os recursos de descrição do SIFT foram aplicados em diferentes espaços de cor, originando as técnicas HSV-SIFT [54] e HueSIFT [55]. Essas técnicas foram comparadas por Van de Sande et al. [56] e foi demonstrado que combinar as características de descrição do SIFT com os espaços de cores é uma maneira promissora de melhorar a precisão no reconhecimento de objetos.

Como pode ser visto, os descritores de ponto flutuante são criados em torno de informações de gradiente local e o objetivo da área é manter o equilíbrio em torno da robustez e velocidade de processamento, preservando a descrição em torno dos gradientes locais. Com essa motivação surgiram os descritores DAISY [57] e HSOG [58].

O descritor DAISY apresenta como conceito central utilizar filtros Gaussianos em múltiplas escalas para convolver vários mapas de gradiente da imagem original.

Uma vez criados os mapas de gradiente convolvidos com os filtros Gaussianos, é criado o descritor pela concatenação de histogramas de orientação seguindo um padrão de janela circular. Umas das principais diferenças trazidas pelo DAISY está no padrão cicular (Figura 2.21) semelhante a uma flor. Nesse padrão o *pixel* central é o ponto de interesse e os outros pontos são as amostras onde as Gaussianas são centralizadas na convolução, cujo desvio padrão é proporcional ao tamanho do raio do círculo delimitador da sub-região. De cada círculo é extraído um histograma de orientação e esses histogramas são concatenados para compor o descritor final. Essas modicações no cálculo do gradiente trouxeram melhorias significativas nos resultados, criando um descritor comparável ao SIFT e até 12 vezes mais rápido.

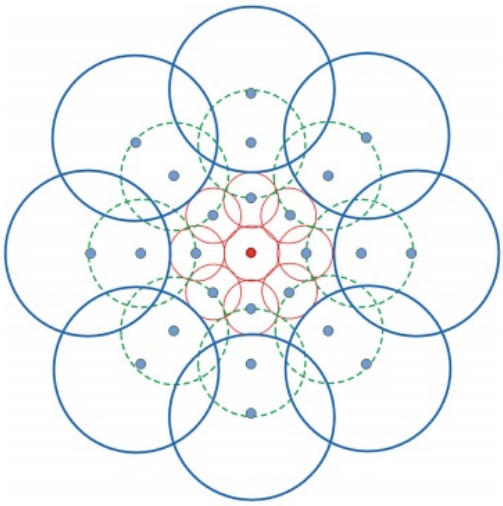


Figura 2.21: (a) O padrão de amostragem usado no DAYSI [6].

O HSOG [58] é um descritor que leva em consideração tanto as informações de gradiente de primeira ordem como de segunda ordem. Segundo os autores as informações de segunda ordem do gradiente carregam informações importantes a respeito da aparência dos objetos. Para a construção do descritor são realizadas algumas etapas: (i) o cálculo de mapas de gradiente de primeira ordem orientados (OGMs); (ii) cálculo de gradientes de segunda ordem com base nos OGMs; (iii) agrupamento espacial; e (iv) redução de dimensionalidade.

Na primeira etapa do HSOG são construídos mapas com as normas do gradiente de cada região em torno dos pontos de interesse da imagem de entrada, de modo que cada mapa descreve o gradiente em um determinada direção. Posteriormente, as direções dos OGMs são quantizadas e são utilizadas para calcular gradientes de segunda ordem das mesmas regiões. Em seguida, é definido um padrão de amostragem para o descritor, já que a a forma do padrão de amostragem implica em várias modificações no desempenho do descritor e a maioria das pesquisas em descritores consiste justamente de alterar o formato e organização das sub-regiões em que

será extraída a informação de gradiente. No HSOG é utilizado o mesmo formato de região de amostragem definido pelo DAISY, pois segundo testes experimentais é o formato de região de amostragem com melhores resultados [57][6]. Definido o formato da região de amostragem e suas sub-regiões, os histogramas de orientação são empilhados um após o outro e é realizada uma etapa de redução de dimensão utilizando a técnica de Análise de Componentes Principais (PCA).

Descritores Binários

Os descritores de ponto flutuante são altamente discriminativos e foram utilizados em muitas aplicações na Visão Computacional no decorrer dos últimos anos. No entanto, o crescimento de aplicações em tempo real e o uso crescente de dispositivos embarcados com pouca memória e com poucos recursos computacionais impulsionaram o desenvolvimento de descritores mais compactos e mais rápidos.

As primeiras tentantivas de tornar os descritores compactos surgiram com a aplicação de técnicas como PCA para redução de dimensão, produzindo resultados significativos [52]. A criação do SURF também impulsionou melhorias, mas o SURF ainda é um descritor de ponto flutuante de 64 posições, ocupando 256 *bytes* de memória por vetor descritor. Em uma aplicação de tempo real em que existem centenas de pontos isso se torna inviável.

Uma outra forma de melhorar o desempenho dos descritores é quantizar suas coordenadas de ponto flutuante em inteiros codificados em menos bits. Em [59][60][61], é mostrado que o descritor SIFT pode ser quantizado utilizando apenas 4 bits por coordenada. Apesar da quantização ser uma operação simples e possibilitar ganhos de memória e de velocidade de correspondência significativos, isso não impede que os descritores de ponto flutuante sejam construídos primeiro, consumindo tempo de processamento.

Uma forma ainda mais radical de encurtar um descritor é binarizá-lo. Em [62] foi utilizada a *Locality Sensitive Hashing* (LSH) [63] para transformar vetores de ponto flutuante em cadeias binárias. No entanto, a binarização, assim como as outras técnicas de encurtamento, precisam que um descritor longo de ponto flutuante seja criado anteriormente, e encurtá-lo envolve também uma quantidade substancial de tempo.

Diante dessas circunstâncias se densenvolveu o grupo de descritores binários que são construídos diretamente a partir de relações de diferenças entre os *pixels* vizinhos ao ponto de interesse seguindo algum padrão de janela.

O primeiro representante da categoria de descritores binários é o BRIEF (*Binary Robust Independent Elementary Features*) [64]. Basicamente, o BRIEF codifica as informações de um *patch* usando apenas simples testes binários, comparando a intensidade entre os *pixels* a partir de um imagem suavizada. Por si só, o BRIEF não é

invariante à escala nem à rotação. No entanto, segundo os autores, seu desempenho é semelhante a um descritor local mais complexo, como o SURF, quando comparado com a sua robustez à iluminação, borramento, e distorção de perspectiva.

O descritor BRIEF é representado por uma sequência binária construída concatenando os resultados obtidos pelo seguinte teste:

$$b = \begin{cases} 1, & S(p_j) > S(p_i) \\ 0, & S(p_j) \leq S(p_i) \end{cases} \tag{2.9}$$

onde $S(p_i)$ representa o valor de intensidade do *pixel* no ponto p_i.

Com relação ao padrão de escolha da ordem de *pixels* o BRIEF utiliza três alternativas. A primeira e mais simples é escolher os pontos de acordo com uma distribuição uniforme, ou seja, os pontos de borda do *patch* apresentam o mesmo peso de serem escolhidos que os mais próximos do centro. A segunda alternativa consiste em utilizar uma distribuição Gaussiana para a escolha, para que os *pixels* do centro, que estão mais livres de distorção tenham maior peso no sorteio de pontos de amostragem. A terceira alternativa consiste em realizar o sorteio de pontos centrando mais de uma Guassiana sobre os pontos, de modo que o resultado pode ser mais polarizado. Como os resultados implicam em poucas diferenças entre as abordagens o sorteio utilizando a distribuição uniforme é o mais utilizado na literatura [64].

Uma vez construído o descritor, o BRIEF introduziu também a utilização da distância de Hamming para a realização da correspondência, que é o equivalente ao somatório do resultado da *string* binária resultante de uma operação XOR entre dois descritores binários. Quanto menor o valor da distância de Hamming mais semelhantes são os pontos.

O BRIEF inicialmente também não foi construído com invariância à escala e rotação. Essas características foram desenvolvidas posteriormente por outras abordagens baseadas nesse descritor.

O descritor ORB (*Oriented Fast and Rotated BRIEF*) é considerado uma dessas abordagens, por ser uma versão invariante a rotação do BRIEF e parcialmente invariante à escala. A invariância a rotação é obtida estimando a rotação do *patch* usando a intensidade do centróide, no qual supera abordagens baseadas em gradientes. Em seguida, o descritor é construído da mesma maneira que o BRIEF. Finalmente, um subconjunto de testes binários é escolhido aumentando o poder de discriminação do algoritmo, que funciona com uma abordagem gulosa selecionando os pares com maior variância, parando quando 256 testes binários são selecionados, implicando na dimensionalidade final do descritor.

Dada a invariância parcial do ORB à escala, surge o BRISK [65], que é um

detector e descritor. Na etapa de detecção o BRISK utiliza muitas características do FAST [66], e na etapa de descrição utiliza características semelhantes ao ORB com algumas modificações.

Na primeira etapa de construção do descritor o BRISK utiliza 60 pontos de amostragem, seguindo um padrão de janelas circulares organizadas de forma concêntrica ao redor do ponto de interesse, como ilustrado pela Figura 2.22.

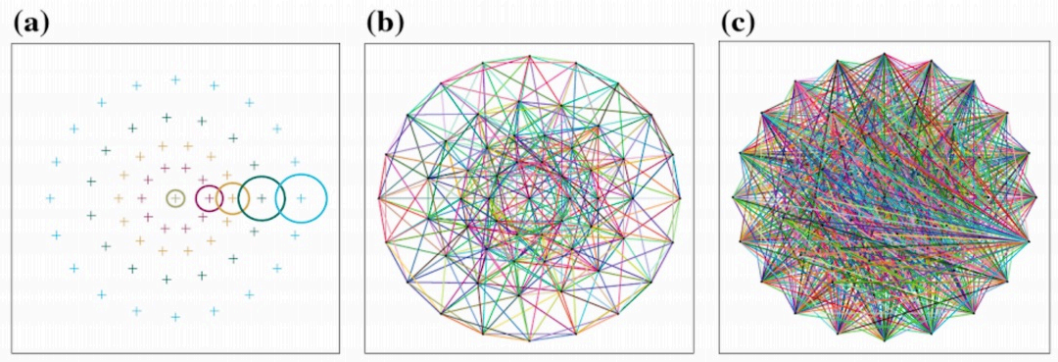


Figura 2.22: (a) O padrão de amostragem usado no BRISK, 60 pontos de amostragem incluindo o ponto central regularmente distribuído em quatro círculos concêntricos ao redor do ponto de interesse. (b) Os pares de curta distância dos pontos de amostra utilizados na construção do descritor. (c) Os pares de longa distância utilizados para determinar a orientação (cada cor indica um par).

O desvio de padrão do kernel da gaussiana de cada pixel de amostragem é definido de acordo com sua distância para o ponto de interesse. Sendo 60 pontos de amostragem $N = 60$ (Figura 2.22 em (a)). Com relação a essas distâncias, dois subconjuntos podem ser obtidos, os pares de curta distância C (Figura 2.22 em (b)) e pares de longa distância L (Figura 2.22 em (c)), matematicamente representados por:

$$C = \{(p_i, p_j)\epsilon I | \, \|p_j - p_i\| < \delta_{max}\} \tag{2.10}$$

$$L = \{(p_i, p_j)\epsilon I | \, \|p_j - p_i\| > \delta_{min}\} \tag{2.11}$$

em que I é a imagem, p_j e p_i são os pontos de amostragem para construir o descritor, δ_{max} é um limiar com valor $9{,}78\delta$ e δ_{min} é outro limiar com valor $13{,}67\delta$, e δ é a escala do ponto de interesse.

Com os dados dos pares pertencentes ao grupo L é definida a orientação do ponto a partir da orientação da média dos gradientes locais obtidos pelos pares de longa distância, dada por:

$$g = \begin{pmatrix} g_x \\ g_y \end{pmatrix} = \frac{1}{||L||} \sum_{(p_i,p_j)\epsilon L} g(p_i, p_j) \tag{2.12}$$

$$\theta = atan2(g_y, g_x) \tag{2.13}$$

em que, (g_x, g_y) são o gradiente médio na direção x e y dos pares de pontos de longa distância e θ é a orientação do ponto-chave.

Utilizando a escala e a orientação do ponto de interesse, o descritor BRISK é definido pela comparação de intensidade, assim como os outros descritores binários. No entanto, apesar das melhorias, principalmente no que diz respeito ao custo computacional, o BRISK apresenta uma qualidade inferior ao desempenho dos descritores SIFT e SURF, mas ainda é considerado uma das melhores alternativas de descritor para aplicações em tempo real [65].

Na tentativa de estender o descritor BRISK e melhorar suas fragilidades, Ortiz propôs um descritor denominado FREAK (*Fast Retina Keypoint*) [67]. Assim como o descritor BRISK, o FREAK tem seu padrão de amostragem baseado em gaussianas, mas com uma distribuição de amostragem de pontos que é biologicamente inspirada no padrão da retina do olho humano.

Basicamente o FREAK apresenta duas diferenças importantes com relação ao BRISK: Consiste em uma alocação de distribuições concêntricas com um crescimento exponencial em relação à distância do ponto-chave. Além disso, o padrão de amostragem cria sobreposições sobre diferentes círculos concêntricos, como ilustrado na Figura 2.23. Essa sobreposição entre as regiões acrescenta redundância e aumenta o poder discriminativo do descritor.

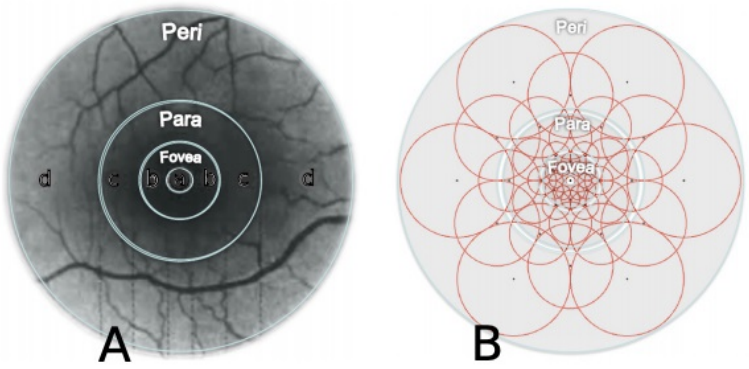


Figura 2.23: Padrões de amostragem do FREAK. Em (a), regiões da retina do olho humano, onde a região ***peri*** (perifoveal) tem um borramento maior, e a área foveal tem um borramento menor. Em (b) as regiões de (a) são representadas por círculos concêntricos de tamanhos diferentes, conforme distribuição Gaussiana.

Como ilustrado pela Figura 2.23 cada círculo representa uma suavização gaussiana com desvio padrão proporcional ao seu tamanho, com a finalidade de reduzir os ruídos na imagem. O centro de cada círculo corresponde a um ponto de amostragem. Na medida em que se aproxima ao centro do padrão, a densidade de pontos de amostragem aumenta, pois este padrão é inspirado na retina do olho humano, no qual a distribuição das células ganglionares, responsáveis pela transmissão da informação da luz ao sistema nervoso, aumenta na medida em que aproxima do centro da retina.

Para o cálculo do descritor, o padrão de amostragem é posicionado no centro de cada ponto-chave. A escala do padrão é ajustada de forma proporcional à escala na qual o ponto-chave foi detectado. Então, seja um par de pontos *(a, b)* do padrão de amostragem, o descritor é definido pela seguinte sequência de bits:

$$D = \sum_{1 \leq i \leq n} 2^{i-1} \tau(a_i, b_i) \tag{2.14}$$

em que n consiste na quantidade de pares de pontos e τ consiste no teste binário dado por:

$$\tau(a, b) = \begin{cases} 1, & I(a) > I(b) \\ 0, & I(a) \leq I(b)) \end{cases} \tag{2.15}$$

em que *I(a)* corresponde a intensidade do ponto a e *I(b)* corresponde a intensidade do ponto b.

Conhecidos os pares de pontos de amostragem que compõem o descritor, o FREAK calcula a orientação do ponto-chave por meio dos gradientes locais utilizando os pares de pontos selecionados. O cálculo da orientação é dado por:

$$O = \frac{1}{n} \sum_{0 \leq i \leq n} (I(a) - I(b)) \frac{a - b}{\|a - b\|} \tag{2.16}$$

em que n corresponde à quantidade de pares utilizados, *I(a)* consiste na intensidade do ponto a e *I(b)* consiste na intensidade do ponto b. Esta orientação é utilizada para rotacionar o padrão sobre o ponto-chave detectado.

Testes realizados em [67] mostram que o FREAK apresentou um desempenho por vezes superior comparado aos descritores BRISK, SURF e SIFT, além de uma redução ainda maior no tempo de processamento. A partir de então, as pesquisas relacionadas a descritores binários buscaram criar um compromisso entre velocidade e desempenho, o que algumas vezes não ocorre, já que os descritores de ponto flu-

tuante ainda são tidos como mais robustos. No entanto, evoluções foram sendo sugeridas por novas abordagens mais recentes.

Em [68] foi proposto o BinBoost que tem como objetivo gerar um descritor binário extremamente compacto e altamente discriminativo, sendo robusto a mudanças de iluminação e pontos de vista. Diferente dos descritores binários mencionados até o momento, que calculam o descritor final com base em simples testes binários comparando a intensidade de *pixels*, cada *bit* gerado pelo *BinBoost* é calculado usando uma função Hash binária da mesma forma como cada *bit* é gerado pelo classificador AdaBoost [69]. Essa função é baseada em *weak learners* que levam em consideração orientações de gradientes de intensidade sobre o *patch* a ser descrito. A função Hash é otimizada de forma iterativa, ou seja, a cada iteração, amostras incorretas serão atribuídas a um peso maior, enquanto o peso das amostras corretas será diminuído. Desta maneira, o próximo *bit* a ser calculado tenderá a corrigir o erro de seus antecessores.

Seguindo a abordagem de binarização após a criação de um descritor de ponto flutuante surgiu o BRIGHT [70]. Nele é criado um histograma de gradientes orientados de um *patch* local centrado em torno de pontos-chave detectados a partir de uma imagem. Os elementos do histograma são então binarizados, e o subconjunto de *bits* é progressivamente selecionado, formando um descritor que varia de 32 a 150 bits. Essa abordagem se mostrou bastante invariante a mudanças de iluminação e oclusões.

Na tentativa de manter um desempenho superior ou igual ao SIFT, com poucos acessos a memória e velocidade baixa, em [71] é apresentado o descritor LATCH, que propõe uma abordagem mais robusta para comparar os valores de pares de *pixels*, exigindo uma reformulação de como os descritores binários são produzidos.

Além dessas abordagens o crescimento da área de aprendizado de máquina proporcionou o surgimento de técnicas utilizando as redes neurais convolucionais (CNNs). O descritor DeepBit [72] é um dos principais dessa área, ele aprende descritores binários compactos de maneira não supervisionada. Os parâmetros dessa rede são atualizados usando *back-propagation* com três critérios: (i) minimizar o erro, (ii) reduzir a correlação de *bits* e (iii) distribuir uniformemente os códigos binários. Seguindo essa linha há também o DELFT [73], em que a descrição do *patch* é treinada por redes neurais convolucionais triplas. Apesar dos bons resultados e de ser uma abordagem recente, as técnicas de descritor baseadas em CNN são computacionalmente complexas para uso em abordagens de tempo real e seu desempenho é comparável ao descritor SIFT, por exemplo.

Como pôde ser visto nesta seção, a tendência das pesquisas utilizando descritores binários é de produzir descritores cada vez mais compactos, que sejam robustos e com menor tempo de processamento. Além disso, tem sido de interesse dos pesquisadores

investigar quais descritores possuem um desempenho melhor para determinadas tarefas. Uma avaliação quantitativa e qualitativa mais detalhada de vários descritores binários pode ser encontrada em [74].

Capítulo 3

Trabalhos Relacionados

De um modo geral, a atenção visual computacional tem sido bastante explorada no decorrer dos últimos anos. As pesquisas nessa área têm como objetivo principal produzir mapas de saliência que tenham uma alta similaridade com os mapas de fixação produzidos pelo rastreamento ocular da atenção humana. De acordo com os resultados obtidos em [75] alguns modelos de atenção já possuem resultados bastante significativos e próximos dos mapas de fixação. No entanto, ainda não existe uma vasta quantidade de aplicações desenvolvidas utilizando os modelos de atenção visual quando comparados a outras ferramentas clássicas já existentes, tais como a classe de algoritmos detectores e descritores de pontos-chave.

Grande parte das aplicações que existem utilizando mapas de saliência ainda fazem uso de uma grande quantidade de ferramentas clássicas para dar suporte à utilização dos mapas. Um exemplo disso é a aplicação de localização de robôs móveis proposto por Siagian [2]. Essa aplicação utiliza o mapa de saliência clássico proposto por Laurent Itti [3], que tem como função ser uma ferramenta limitadora de regiões de busca para o algoritmo SIFT [35], que é responsável por identificar pontos-chaves de uma região saliente armazenada no banco de imagens do robô.

Nessa aplicação o objetivo é reduzir o tempo de processamento do SIFT aplicando-o apenas nas regiões mais importantes da cena. No entanto, para isso são feitos ajustes com relação ao tamanho e a quantidade de regiões selecionadas pelo mapa de saliência, provocando uma distorção da área delimitada originalmente pelo mapa. O objetivo dessas adaptações é ajustar a saída do mapa para que se torne possível a aplicação do algoritmo SIFT na região selecionada, pois se forem selecionadas regiões muito pequenas e que não possuem bordas abruptas, o SIFT não irá encontrar pontos na etapa de detecção, e posteriormente não existirão pontos na etapa de descrição. Para minimizar esse problema Siagian propôs empiracamente um tamanho fixo de região e uma quantidade de pelo menos 5 regiões de saliência por *frame*, de modo que só pode haver 60% de sobreposição entre elas, senão a região é descartada.

Quando não são feitas adaptações com relação ao tamanho das regiões selecionadas pelo mapa de saliência, a qualidade do resultado de correspondência pode variar bastante a depender da resposta do mapa. Em [76], os mapas de saliência são utilizados como limitadores de região de busca considerando que as regiões passadas para o SURF são as regiões encontradas pelos mapas sem adaptações. Os resultados obtidos mostram que sem o ajuste do tamanho da região, a tarefa de correspondência pode ser totalmente prejudicada impossibilitando muitas vezes o rastreamento de pontos.

Como grande parte dos métodos de detecção de pontos foram desenvolvidos para trabalhar com regiões grandes que possuam bordas e cantos de objetos, algumas aplicações realizam primeiro a detecção de pontos invariantes na imagem completa e em seguida, de forma desacoplada, aplicam um método de detecção de saliência. As regiões selecionadas pelo método de saliência são utilizadas como uma máscara de filtragem para eliminar pontos. Eliminados os pontos que estão fora dessa máscara, apenas os pontos dentro das regiões consideradas mais importantes são utilizados nas etapas de construção do descritor e de *matching*.

O trabalho proposto em [66] realiza esse processamento desacoplado na tarefa de recuperação de conteúdo de imagens. Nesse trabalho o SURF é utilizado como detector e descritor de pontos e o sistema de atenção visual VOCUS [77] é utilizado para delimitar a máscara de seleção de pontos. Além disso, a informação de saliência é utilizada para estimar a similaridade entre os dois pares da imagem. Os resultados experimentais demonstraram um desempenho atraente desse novo método. Ocorreu uma redução significativa de pontos na etapa de correspondência de uma ordem de grandeza com uma perda de desempenho próxima a 10%. Apesar dos ganhos significativos nas etapas de correspondência e *matching*, essa abordagem tem como ponto fraco o custo adicional da construção do mapa em conjunto com a detecção custosa na imagem completa, portanto sendo mais viável para aplicações em que as operações de correspondência e *matching* são críticas e o tempo de detecção e filtragem não são tão relevantes.

Em [78] é utilizado um modelo de atenção para classificar todos os pontos-chave encontrados pelo SIFT de acordo com o seu nível de atenção, e somente os pontos mais distintos são utilizados. Para cada imagem, após a extração do SIFT o modelo de atenção é usado para gerar o mapa de saliência, e então um método de crescimento difuso é executado para selecionar todas as regiões salientes. Por conta do custo computacional o número de regiões é limitado a 3. As regiões são delimitadas por retângulos e os pontos são ponderados pelo valor de saliência e pela distância dos pontos ao centro da região de saliência. Selecionados os pontos dessa forma, são construídos os descritores e os pontos são utilizados na recuperação de imagens em bases de dados. Os resultados dessa aplicação mostram que centenas de pontos

desnecessários de fundo encontrados pelo SIFT são eliminados pelo sistema atencional, provocando uma melhoria significativa na precisão da recuperação mesmo com o aumento do tamanho da base de teste. No entanto, é preciso processar a etapa de detecção de pontos de forma desacoplada à etapa de construção de saliência e isso aumenta o tempo da etapa de detecção.

Em [79], o sistema de atenção visual proposto por Itti é utilizado em conjunto com o SIFT para o reconhecimento de objetos. O processamento do mapa e da detecção de pontos também são realizados de forma desacoplada. Além disso, é utilizado o detector de bordas Canny para encontrar a delimitação correta dos objetos, já que o mapa de saliência pode suprimir essa delimitação facilmente. Caso uma região seja delimitada pelas arestas dos objetos encontrados pelo algoritmo de Canny e dentro dessa região existam áreas importantes delimitadas pelo mapa de saliência, apenas os pontos encontrados nessas áreas serão considerados nas etapas de descrição e *matching*. Os resultados obtidos para esse tipo de abordagem mostraram uma pequena melhoria na performance de recuperação das imagens, assim como uma redução significativa de pontos detectados. No entanto, segundo os autores, utilizar um algoritmo extra para delimitar as regiões dos objetos não é uma característica desejável para o sistema e diferentes modos de selecionar essas regiões causam impactos diferentes no sistema de recuperação.

No trabalho proposto por Lopez em [80] o mapa de saliência WMAP é avaliado como limitador de espaço de busca em uma aplicação de navegação de um robô móvel. Nesse trabalho são utilizados os algoritmos SIFT e SURF e é feito um estudo comparativo dos dois quando utilizados em conjunto com o mapa de saliência. Nessa aplicação o processamento dos pontos do detector e a construção do mapa também são realizados de modo desacoplado, ficando sob função do SIFT e SURF reconhecer com o auxílio de uma ferramenta de classificação uma cena já visitada pelo robô e a partir daí extrair informações importantes para a localização dele no ambiente. Nos experimentos realizados as regiões foram delimitadas de acordo com a variação do limiar de corte realizado manualmente. Foram utilizados 7 limiares com valores entre 0 e 1 para delimitar as regiões de saliência.

Nos experimentos realizados nesse trabalho utilizando somente o SURF, resultados ruins foram obtidos, mas que por vezes os resultados foram melhorados ou piorados de acordo com o limiar de saliência selecionado, o que demonstra a falta de correlação existente entre as áreas de saliência e os pontos detectados pelo algoritmo SURF, demonstrando que o uso do mapa pode afetar drasticamente o desempenho dos detectores, mesmo reduzindo o tempo de processamento. Já no uso com o SIFT o desempenho só foi de fato afetado em cerca de 50% quando o limiar de saliência foi muito alto (acima de 0.8), o que demonstra que o mapa, quando utilizado com o SIFT, apresentou resultados melhores que quando utilizado com o SURF. No en-

tanto, para os dois casos, ocorreu uma queda significativa de tempo de busca nas bases do robô.

Grande parte das aplicações realiza o processamento desacoplado do mapa de saliência e do detector de pontos, mas isso gera tempo de processamento extra, além de incluir uma etapa de filtragem posterior, o que muitas vezes não é desejável. O trabalho proposto por Mesquita [81] de reconhecimento de instâncias realiza adaptações em um mapa de saliência para que ele selecione objetos e delimite bem suas bordas. Construído o mapa de saliência, a imagem é dividida em *patches* de modo que cada um é ponderado com o valor médio de saliência. A partir de então se inicia a busca visual pelo elemento que se deseja encontrar na cena. A busca atua primeiro nos *patches* de maior saliência, sendo que nesse momento o detector SURF é utilizado para realizar a extração de pontos de interesse. Esses pontos são utilizados posteriomente na etapa de reconhecimento do elemento que se deseja encontrar. Vale ressaltar que foram realizadas modificações no SURF para que seu comportamento seja adequado ao tamanho do *patch* em que é utilizado, ou seja, o SURF deixa de ser uma ferramenta de extração de características global e se torna uma ferramenta local, chamada de *patch-based SURF*. Além disso, a alocação de recursos do *patch-based SURF* é controlada pela importância do *patch* de saliência.

A vantagem dessa abordagem é a melhoria no tempo de reconhecimento dos objetos. Os experimentos demonstram que ocorreu uma redução de até 80% do tempo de processamento em alguns casos de teste, quando a abordagem é comparada com a busca clássica. No entanto, essa abordagem, assim como as outras encontradas na literatura, provocam distorções na saída do mapa de saliência com o objetivo de selecionar as bordas dos objetos, que guardam as informações buscadas pela maior parte dos detectores de pontos utilizados na visão computacional. Isso provoca uma distorção do que de fato é encontrado pelos algoritmos baseados na atenção visual humana.

Existem abordagens que realizam modificações ainda maiores, chegando a modificar etapas de construção dos mapas de saliência para extrair pontos diretamente do mapa. O modelo de atenção NLOOK desenvolvido em [82] tem esse objetivo. Ele é uma modificação do modelo clássico NVT. A principal modificação realizada diz respeito à troca das etapas de subtração dos mapas em diferentes escalas (etapa de centro-periferia) pela construção do espaço-escala utilizando o operador Laplaciano, que é de fato o que é feito pelo algoritmo SIFT, com a diferença que esse processo é realizado para cada canal de característica. Além dessa modificação o mapa de saliência final não é construído em uma escala menor que a imagem original, que é o que ocorre na maioria das abordagens de modelos de atenção, mas sim na mesma escala da imagem original.

Essas modificações geram maior estabilidade do mapa de saliência com relação às

principais transformações em imagens, já que o operador utilizado em sua construção é o Laplaciano. No entanto, as respostas obtidas pelo mapa são diferentes das respostas obtidas pelo modelo NVT, e não há validação se do ponto de vista da atenção visual o modelo criado tem uma proximidade considerável com os mapas de fixação, ou se a sua resposta está mais próxima de um detector clássico. Além disso, já seria esperada a melhoria desse modelo com relação às transformações e sua possível extração de pontos, já que o operador Laplaciano é bastante conhecido e utilizado na literatura para essa finalidade.

Não há muitos indícios de que mapas de saliência sejam de fato utilizados além de uma ferramenta de seleção e segmentação de regiões, assim como não há trabalhos investigando a possibilidade de utilizá-los além dessa funcionalidade dentro da visão computacional. Em [36] é discutido que o objetivo inicial dos modelos de atenção é puramente teórico, apenas para tentar replicar alguns aspectos do sistema visual humano, mas que a área inspirou o desenvolvimento de algumas abordagens de detecção de pontos. O detector proposto por Kadir [46] e em seguida aprimorado em [47] é inspirado nos conceitos de atenção humana e utilizado em algumas aplicações de SLAM para detectar o fechamento de loops [83]. No entanto, vale ressaltar que esses detectores desenvolvidos com inspiração no conceito de saliência não constroem mapas de saliência, apenas utilizam os conceitos como inspiração para propor técnicas de extração de pontos utilizando o conteúdo da imagem. O detector de Kadir, por exemplo, faz a extração buscando máximos de entropia em *patches* da imagem de entrada em escala de cinza.

Apesar da ausência de discussões a respeito do uso de mapas de saliência como detectores de pontos e dos mapas possuírem tais características, em [84] um mapa de saliência é avaliado com relação à sua repetibilidade de pontos em alguns pares de cenas sob transformações simples. Esse trabalho ainda analisa o problema de um modo pouco aprofundado, sendo que não são realizados muitos testes, nem são avaliados diferentes modelos de atenção visual. Além disso, a base de testes é bastante simples e em todas as imagens sempre existe um elemento nitidamente saliente na cena, o que facilita o trabalho do sistema de atenção, já que é mais dificil para o sistema lidar com ambientes com vários distratores. Os resultados apontam uma boa repetibilidade das regiões encontradas pelo mapa de saliência, mas ainda são necessárias análises bem mais aprofundadas para garantir a viabilidade de modelos de atenção visual como ferramentas para detecção de pontos.

Capítulo 4

Investigação em Modelos Computacionais de Atenção

Esse capítulo descreve as investigações que foram realizadas nos modelos computacionais de atenção visual e os critérios utilizados para verificar a possibilidade desses modelos serem utilizados como extratores de características locais.

A extração de características é composta de duas fases: a fase de detecção e a fase de descrição, sendo que, a fase de descrição é de certa forma dependente da detecção, pois o descritor só pode ser feito quando se conhece as características da vizinhança dos pontos que foram detectados. Além disso, como descrito no Capítulo 2, na Seção 2.4, os pontos detectados precisam satisfazer alguns requisitos básicos, sendo o principal deles a invariância, pelo menos com relação às transformações mais básicas de uma imagem.

Em um mapa de saliência há muitos pontos conectados, já que eles formam regiões. Como os mapas são muito densos com relação à quantidade de pontos, não é vantajoso considerar todos os pontos como pontos-chave. Até mesmo porque nem todos os pontos que compõem a região de saliência são realmente invariantes. No entanto, verificar e desenvolver uma estratégia de seleção de pontos salientes que sejam invariantes às transformações não é uma tarefa simples. Vale ressaltar que nesse trabalho se deseja utilizar apenas o conteúdo de saliência para selecionar os pontos, como se o mapa de saliência fosse um filtro de extração de características.

O primeiro motivo de não ser uma tarefa simples se trata da variedade de abordagens computacionais existentes com respostas bastante diferentes, como ilustrado na Figura 4.1. Cada abordagem é feita com uma ferramenta matemática diferente, e isso influencia muito a questão de invariância, pois algumas ferramentas matemáticas não foram feitas para suportar deformações na imagem. Além disso, o mapa de saliência não é uma resposta de apenas um filtro aplicado na imagem de entrada, mas uma junção de respostas de vários filtros e de várias transformações que são realizadas na imagem.

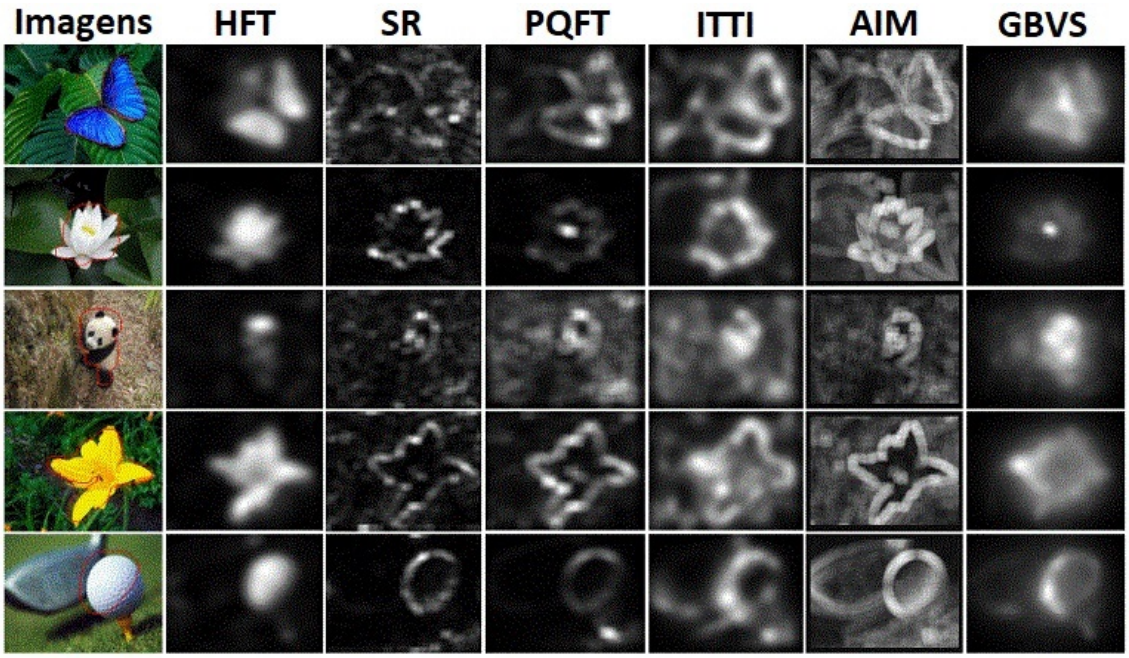


Figura 4.1: Exemplo de algumas abordagens de mapas de saliência com respostas bastante distintas.

Para contornar o problema da grande variedade de abordagens existente, esse trabalho fez uma filtragem de abordagens, elimando algumas da investigação. O pré-requisito da filtragem é a similaridade com os mapas de densidade das fixações. Já que o objetivo é investigar o potencial de sistemas atencionais na extração local de características, não faz sentido considerar a análise utilizando abordagens que possuem resultados que se afastem da resposta dos mapas de densidade. Na Figura 4.1, por exemplo, as abordagens SR, PQFT e AIM se afastam bastante de mapas de densidade e são mais semelhantes a detectores de borda.

Quanto mais o mapa de saliência se afasta do mapa de densidade, mais ele é parecido com um detector de bordas. Nestes modelos existe mais distinção entre os pontos e mais gradiente local, parecendo ser uma tarefa mais simples de selecionar pontos invariantes. No entanto, quanto mais próximo do mapa de densidade, os pontos dos mapas perdem o grau de distinção entre si, o gradiente local se torna menos notável e é mais difícil adotar um critério de seleção de pontos. Mesmo assim, os modelos escolhidos para a investigação nesse trabalho são os mais próximos dos mapas de densidade, pois eles refletem uma proximidade maior com os mapas que são considerados referência na área de atenção visual. Vale ressaltar que essa medida de proximidade é dada por um conjunto de métricas clássicas muito utilizadas na literatura para medir a similaridade entre mapas de saliência e mapas de densidade das fixações, essas métricas podem ser encontradas em [14] [32].

Foram selecionados os modelos de atenção que refletem melhor o comportamento do mapa de densidade e que estão disponíveis para testes. Já os modelos que são

mais semelhantes a detectores de borda foram eliminados dessa investigação. Os modelos utilizados foram: SAM-ResNet, SAM-VGG, LDS, GBVS e o modelo clássico proposto por Itti.

Os modelos SAM-ResNet e SAM-VGG quantitativamente são os mais próximos dos mapas de densidade (apresentam similaridade média em torno de 70%, o máximo atingido na área até o momento). Eles são construídos utilizando técnicas de Deep Learning. Há várias abordagens construídas utilizando essas técnicas, todas com resultados de similaridade muito semelhantes, com diferenças de apenas 1%. Como todas essas abordagens têm respostas bastante semelhantes, foram escolhidas duas que têm uma representação significativa na área para representar os modelos de atenção baseados em aprendizado de máquina.

Com relação aos modelos que não utilizam Deep Learning, as técnicas LDS e GBVS têm uma similaridade razoável, mas com algumas diferenças com relação aos mapas de densidade. Apesar das diferenças significativas com os mapas de densidade, as técnicas não geram respostas nem um pouco semelhantes a detectores de borda, e dentro do grau de similaridade e de limitação de ferramentas em que foram construídas, elas representam bem a categoria de mapas que são criados sem técnicas de aprendizado de máquina.

Por fim, foi incluído o modelo clássico proposto por Itti que, apesar de não ser tão próximo dos mapas de densidade, é o primeiro modelo computacional de atenção criado na área, então foi considerado importante incluir esse modelo na investigação. Além disso, a resposta desse modelo também é bastante distante de detectores de borda. A Figura 4.2 ilustra um exemplo dos mapas de cada modelo que foram considerados nessa investigação.

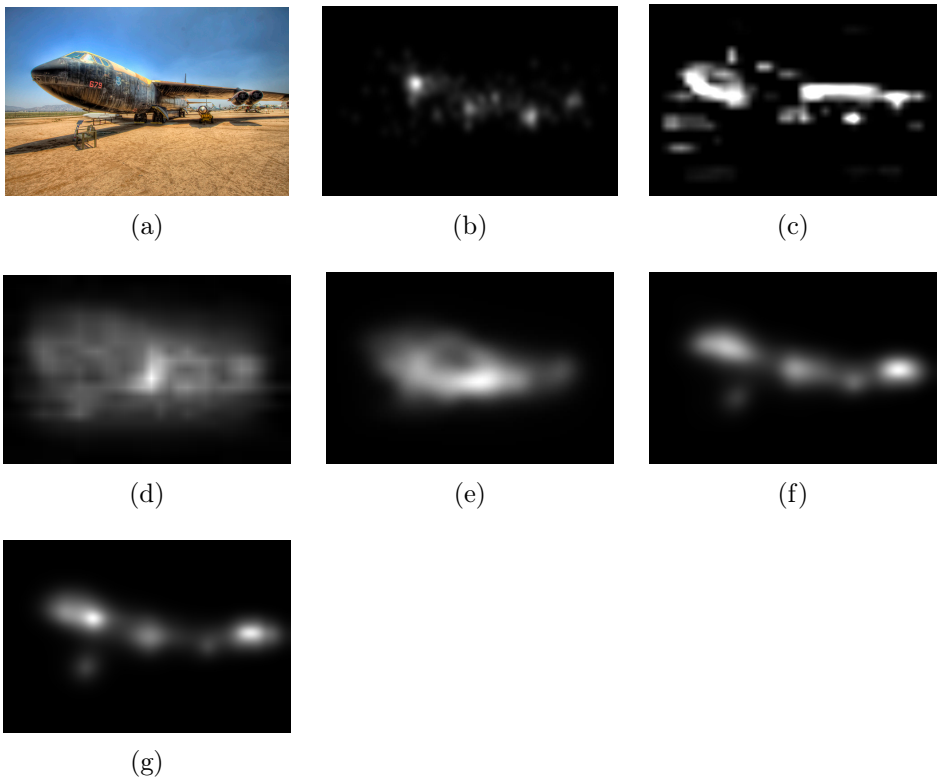


Figura 4.2: Exemplos de mapas de saliência das abordagens utilizadas nesse trabalho. (a) Imagem de entrada. (b) Mapa de densidade das fixações. (c) Mapa de saliência clássico. (d) Mapa de saliência GBVS. (e) Mapa de saliência LDS. (f) Mapa de saliência SAM-VGG, (g) Mapa de saliência SAM-ResNet.

Definido o critério de seleção das abordagens que foram utilizadas na investigação, o principal desafio está em estabelecer um critério de seleção de pontos apenas com a informação de saliência, e que esses pontos sejam invariantes às transformações. Para tentar estabelecer um critério de seleção de pontos, os mapas foram perturbados com algumas transformações e foi observado que as regiões de saliência se mantêm semelhantes.

Como ilustrado na Figura 4.3, mesmo após as transformações, as regiões dos mapas se mantêm, mas com modificações perceptíveis. Essas modificações a nível de vizinhança local são bastante significativas em todas as abordagens testadas, tornando extremamente difícil a tarefa de seleção de pontos.

Selecionados os modelos que são utilizados na investigação o próximo passo é definir um método para a seleção de pontos de interesse utilizando os pontos de saliência. O primeiro critério de busca adotado foi o de utilizar o que se tem de mais próximo na área de um sistema de seleção de pontos salientes: a rede de extração de focos atencionais proposta por Itti e Koch em seu primeiro trabalho. Essa rede tenta simular o processo de sacadas humanas quando estamos varrendo visualmente uma determinada cena. Se trata de uma rede *winner-takes-all* cujo objetivo é encontrar

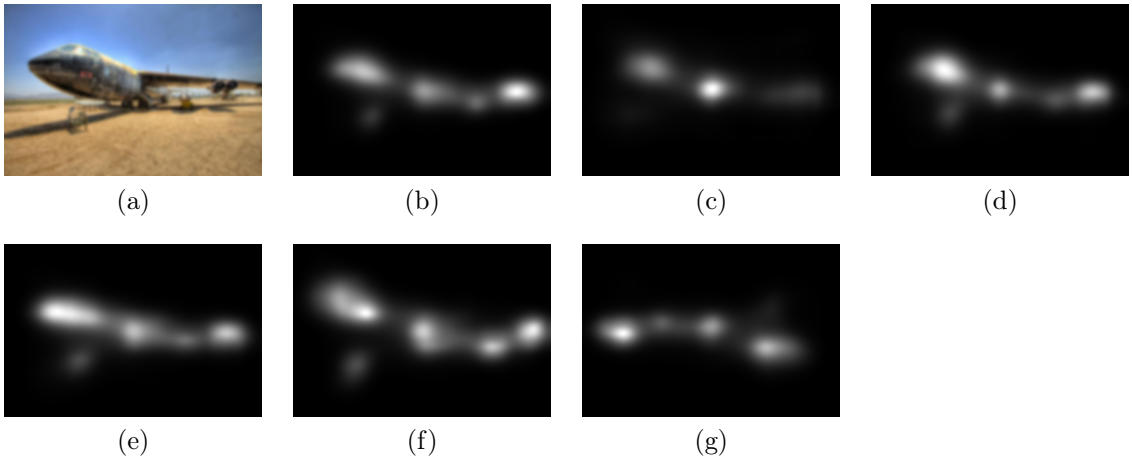


Figura 4.3: Exemplo da resposta do modelo SAM-VGG quando pertubado com imagens transformadas. (a) Imagem Original. (b) Mapa da imagem original. (c) e (d) Mapa das imagens borradas. (e) Mapa da imagem escurecida. (f) Mapa da imagem comprimida. (g) Mapa da imagem rotacionada.

máximos globais no mapa de saliência. Esses máximos representam as áreas de maior estímulo. Esse processo é iterativo, sendo que, uma vez encontrado o máximo global, ele é suprimido junto com a sua vizinhança localizada até uma determinada distância pré-definida, como foi explicado e demonstrado no Capítulo 2, na Seção 2.2.1.

Essa rede foi testada e utilizada para extrair os pontos, porém não foram obtidos bons resultados, pois não há garantias de que os mapas que são construídos a partir de imagens transformadas vão manter a vizinhança dos pontos salientes intacta para que eles continuem sendo os mesmos máximos globais em cada iteração. Além disso, como uma parte da vizinhança do ponto selecionado é apagada em cada iteração e não é utilizada na seleção, muitos pontos salientes que foram salientes na imagem sem transformações podem ser apagados e nunca encontrados no processo de seleção de pontos salientes em imagens transformadas, justamente por sua saliência e sua relação com os pontos vizinhos ter sido alterada, como ilustrado na Figura 4.4. Como pode ser visto nessa figura, há mudanças sutis entre os mapas de saliência e essas mudanças fazem com que pontos diferentes sejam escolhidos em cada imagem. Apesar dos pontos estarem sob as mesmas regiões da imagem, eles não formam pares de correpondência corretamente, pois estão em locais diferentes, deixando o sistema de extração sem robustez, o que não é desejado.

Os resultados também podem ser justificados pelo fato de que quando um ponto é selecionado pela rede, ele é selecionado em baixa dimensão, ou seja, um *patch* é selecionado em baixa dimensão e será representado por um ponto na dimensão da imagem original. Para definir onde será a posição do ponto em alta dimensão há duas estratégias: na primeira, a posição é definida de forma aleatória dentro

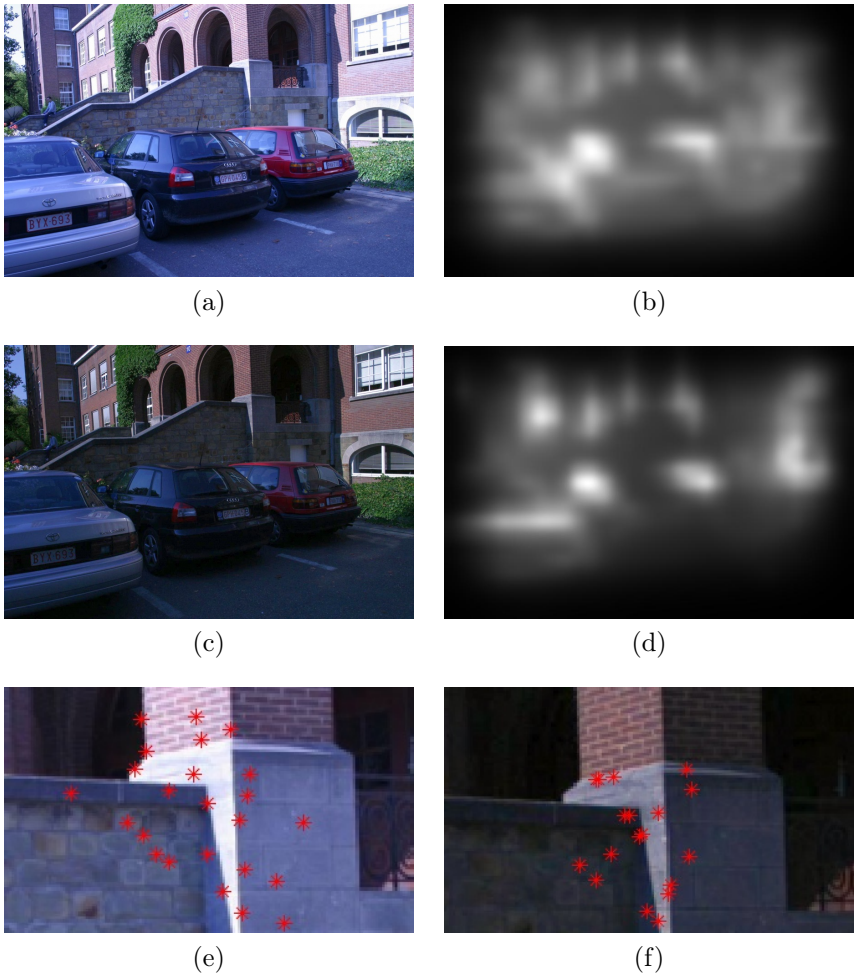


Figura 4.4: Pontos encontrados utilizando a rede neural *winner-takes-all*. (a) Imagem original. (b) Mapa de saliência da imagem original. (c) Imagem com iluminação alterada. (d) Mapa de saliência da imagem com iluminação alterada. (e) Região da imagem original com os pontos encontrados pela rede *winner-takes-all*. (f) Região da imagem com iluminação alterada com os pontos encontrados pela rede *winner-takes-all*.

da janela de *pixels* que representa o *patch* em baixa dimensão, e na segunda o ponto é definido como o *pixel* em alta dimensão que ocupa a posição equivalente ao centro do *patch* em baixa dimensão. Esse detalhe contribui fortemente para resultados ruins na detecção, pois não há garantias do correto posicionamento do ponto utilizando a primeira alternativa. Ao utilizar a segunda estratégia, os pontos são sempre selecionados em janelas quadradas de tamanho fixo, mas com a técnica de supressão de pontos utilizando uma janela de tamanho fixo a seleção ainda pode ocorrer em lugares bem diferentes.

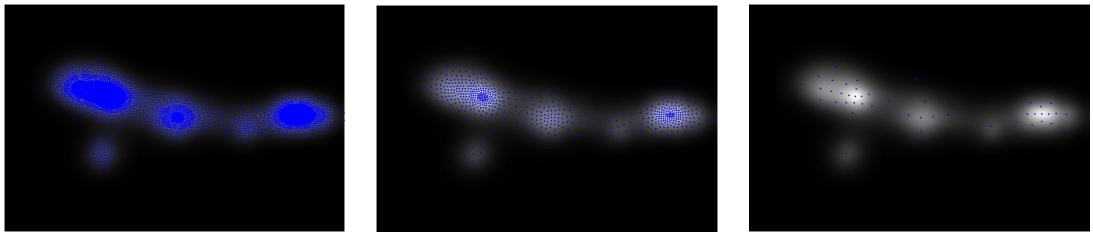
Apesar das fragilidades desse método, foi observado que se ele fosse adaptado para ser utilizado em mapas com a dimensão original da imagem e com um tamanho de janela ajustável o problema de suprimir pontos que poderiam ser salientes poderia ser atenuado. Então, foi criada uma versão extremamente simples desse método e

sem utilizar uma rede neural. Basicamente o método seleciona máximos globais de forma iterativa e o tamanho da região de supressão é ajustado pela seguinte equação:

$$t = \frac{(x - M)}{M} \tag{4.1}$$

em que t é o tamanho da janela de supressão, x é um paramêtro livre que controla a sensibilidade da janela e M é o valor o máximo que foi selecionado.

Esse método proporciona uma seleção que faz mais sentido do ponto de vista da saliência, pois em regiões mais salientes são encontrados mais pontos e em regiões menos salientes menos pontos, como ilustrado pela Figura 4.5. Além disso, essa abordagem tem a fragilidade de precisar de um ajuste de um parâmetro livre, cujo melhor valor pode ser bem diferente de imagem para imagem, e a escolha desse limiar define bastante a qualidade dos resultados.



(a) (b) (c)

Figura 4.5: Pontos encontrados pelo método de seleção de máximos globais. (a) Método utilizando limiar de sensibilidade x = 300 . (b) Método utilizando limiar de sensibilidade x = 800. (c) Método utilizando limiar de sensibilidade x = 3000.

O problema dessa abordagem novamente é que um ponto pode ser selecionado como máximo global em uma imagem e não ser selecionado na imagem transformada no mesmo local, e isso pode acontecer por vários motivos: ou porque o valor de saliência do ponto foi modificado e ele não é mais um máximo global, ou porque ele faz parte de uma região com vários pontos iguais e qualquer um desses pontos pode ser o máximo global, ou porque ele seria um máximo global mas foi suprimido pela janela de supressão antes de ser considerado um máximo global. Além disso, se as regiões de saliência são muito afetadas por uma transformação isso faz com que a região por completo tenha tamanhos de janelas diferentes, proporcionando seleções totalmente diferentes. Gerando resultados ruins, como ilustrado na Figura 4.6. Como observado nessa figura, os pontos encontrados nos pares de regiões I, II e III não são coincidentes, eles estão em pequenas regiões diferentes da imagem, o que não é desejável para um detector.

Em seguida, foi pensado em aprimorar o método separando as regiões que têm o mesmo valor de saliência (platôs) antes de escolher o valor de máximo global, sendo

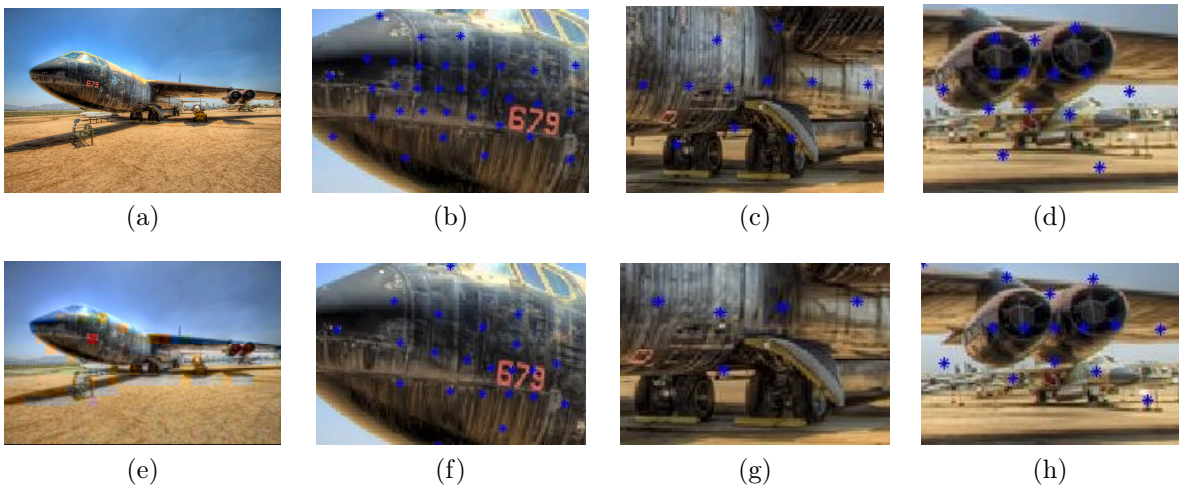


Figura 4.6: Seleção de pontos utilizando o método dos máximos globais com limiar x = 300. (a) Imagem Original. (b) Região I da imagem original com pontos encontrados utilizando o método dos máximos globais. (c) Região II da imagem original com pontos encontrados utilizando o método dos máximos globais. (d) Região III da imagem original com pontos encontrados utilizando o método dos máximos globais. (e) Imagem Comprimida. (f) Região I da imagem comprimida com pontos encontrados utilizando o método dos máximos globais. (g) Região II da imagem comprimida com pontos encontrados utilizando o método dos máximos globais. (h) Região III da imagem comprimido com pontos encontrados utilizando o método dos máximos globais.

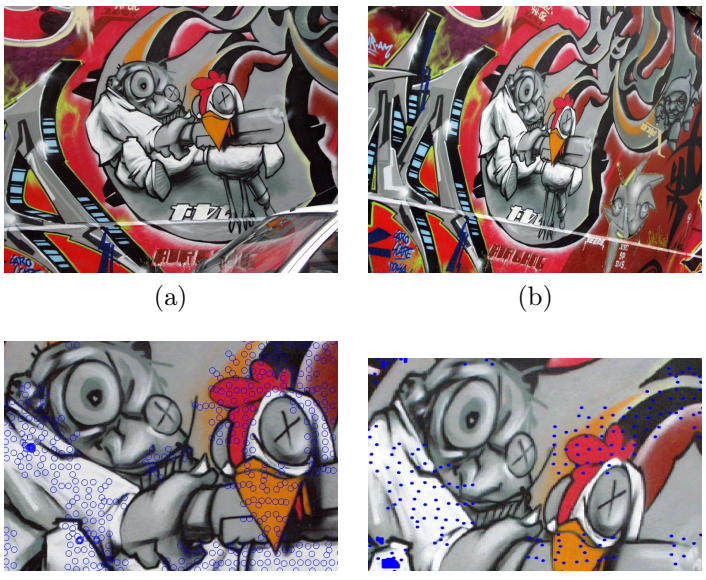
que o valor do máximo global seria o centróide do platô com maior valor de saliência. Além disso, o tamanho da janela seria o tamanho do platô. Essa abordagem também não trouxe bons resultados, pois o tamanho dos platôs é variável de acordo com a transformação e seleções diferentes terminam sendo realizadas novamente.

A partir dessas tentativas, evidenciou-se que tentar buscar máximos globais isolados de forma iterativa não é uma boa estratégia. Então foi tentada uma abordagem de selecionar um ponto-chave a partir da relação que ele tem com a vizinhança e com seus valores de saliência. Para isso, foram tentados algoritmos de agrupamento. Inicialmente, foi utilizado o algoritmo mais simples, o *k-means*. Apesar de parecer resolver problemas das tentativas anteriores, essa técnica traz novos problemas: o primeiro deles é que são muitos pontos de saliência que serão utilizados no agrupamento, o segundo problema é que é preciso definir um número de centróides a *priori*.

Para contornar o problema de muitos pontos, o *k-means* foi aplicado iterativamente e localmente nos *patches* mais salientes do mapa. Para isso, foi feita uma limiariazação do mapa de saliência ainda em baixa dimensão utilizando a algoritmo de Bersen [85] para separar as regiões mais salientes das regiões descartáveis. Separados os *patches* o algoritmo segue de forma iterativa selecionando os máximos globais, um a cada iteração, com o mapa em baixa dimensão. Quando um *patch* é

selecionado, o *k-means* é utilizado para encontrar a posição de uma quantidade de centróides pré-definidos no mapa de saliência em alta dimensão, sendo que esse agrupamento atua somente nas limitações de tamanho do *patch* selecionado em baixa dimensão.

Utilizando essa abordagem foram utilizadas quantidades de centroíde que variaram de 1 até 10, mas não foram observados bons resultados, principalmente porque quando se rotaciona ou se modifica a escala ou quando ocorrem mudanças de perspectiva, o conteúdo do *patch* de saliência se modifica bastante e os centróides de uma imagem não são nem um pouco compatíveis com os da outra. Além disso, a técnica sofre do mesmo problema de todas as tentantivas anteriores, que é a modificação dos valores de saliência e a da relação local entre os pontos quando uma transformação é realizada em uma imagem, apesar de a grosso modo regiões parecidas se repetirem nas imagens transformadas. A Figura 4.7, mostra uma exemplo desse problema, pois nas regiões destacadas os pontos que fazem parte de uma mesma região em cada imagem estão em locais diferentes, tornando esse um método também sem robustez.



(a) (b) (c) (d)

Figura 4.7: Seleção de pontos utilizando o *k-means* localmente. (a) Imagem original. (b) Imagem com mudança de Perspectiva. (c) Pontos selecionados em região da imagem original utilizando o *k-means* de forma iterativa. (d) Pontos selecionados em região da imagem transformada utilizando o *k-means* de forma iterativa.

Baseadas em todas as dificuldades encontradas nas tentativas realizadas utilizando métodos locais de seleção de pontos, chegou-se à conclusão de que a busca por pontos que sejam invariantes dentro de mapas de saliência não é factível ou pode não ser facilmente identificável por um ser humano. Novas abordagens foram pensadas para continuar investigando o problema, mas por uma série de motivos essas idéias foram descartadas também. Esses motivos foram divididos em 5 categorias

de problemas, são eles: o problema do gradiente, problema dos máximos locais e globais, o problema da baixa dimensionalidade, o problema da explosão de amostras e o problema da busca cega. Esses problemas e as técnicas que foram descartadas por conta deles são melhor descritos nas Seções 4.1, 4.2, 4.3, 4.4 e 4.5.

4.1 Problema do Gradiente

Foi avaliada a possibilidade de aplicar filtros que buscam por pontos de maior variação diretamente nos mapas de saliência, tais como o filtro de Kishy [86] e o Laplaciano [86], o problema é que esses filtros detectam pontos em regiões de alta variação, ou seja, alto gradiente. O mapa de saliência tem um gradiente muito suave porque ele é formado em grande parte por regiões de platôs. As áreas que possuem um gradiente mais significativo nos mapas de saliência ficam nos arredores dos platôs e essas áreas são ainda mais instáveis, pois elas podem existir em um mapa, mas podem não existir em outro mapa da imagem que sofreu alguma transformação. Ou seja, os pontos de maior gradiente em uma mapa de saliência têm uma tendência a serem mais instáveis e não representam bem a idéia de saliência, pois a idéia central é que pontos mais salientes deveriam ser utilizados, pois são os mais interessantes, e não os pontos localizados nos extremos das regiões de saliência. Dessa forma os filtros que são conhecidos para extração e seleção de pontos não podes ser utilizados no conteúdo de saliência, deixando pouquíssimas possibilidades de ferramentas de seleção.

4.2 Problema dos Máximos Locais e Globais

Após as tentativas falhas de selecionar iterativamente máximos globais, foi avaliada a possibilidade de adotar estratégias de busca de máximos locais de saliência. Essas estratégias funcionam pela simples comparação de cada ponto com seus 9 vizinhos, se ele tiver um valor maior que todos eles ele é um máximo local, senão ele é descartado e a busca segue para os próximos pontos. Essa estratégia é muito utilizada dentro dos algoritmos de deteção clássicos para realizar a busca de pontos após a aplicação de algum filtro. O SIFT por exemplo, utiliza essa estratégia buscando máximos locais em escala.

O problema da abordagem de busca de máximos locais é que não há máximos locais nos mapas de saliência, as regiões são tipicamente cobertas por platôs, demonstrando a baixa distinção entre os pontos de saliência e sua vizinhança local.

4.3 Problema da Baixa Dimensionalidade

Após as dificuldades encontradas em utilizar gradiente e máximos locais, foram investigados os aspectos de construção dos mapas e foi observada a baixa dimensionalidade em que todos são construídos. Um mapa de saliência na verdade é construído em uma dimensão por vezes mais que 10 vezes menor que a dimensão da imagem de entrada. E nessa construção, as características da imagem em alta dimensão dão origem a um mapa que atua como um verdadeiro redutor de informação. Posteriormente, os mapas são borrados para eliminar a discrepância entre os valores dos *patches* e redimensionados para a dimensão da imagem de entrada. Essa construção explica o motivo da grande quantidade de platôs e da falta de distinção entre os pontos. Toda essa redução de informação causa uma enorme dificuldade em desenvolver uma técnica que consiga lidar com a falta de distinção dos elementos.

4.4 Problema da Explosão de Amostras

A partir do problema da falta de distinção de pontos e da formação de platôs, foi avaliada também a possibilidade de modelar o mapa de saliência como uma superfície composta por um conjunto de gaussianas. A idéia é que essa superfície consiga se ajustar diante das imperfeições da superfície original de saliência, gerando assim superfícies em que possam ser extraídos pontos com algum critério. No entanto, realizar uma tarefa como essa requer utilizar todos os dados da superfície de saliência para o ajuste das curvas. É uma tarefa extremamente lenta computacionalmente. Além disso, é uma tarefa em que a quantidade de gaussianas para representar a função precisa ser definida imagem por imagem, pois uma quantidade pode representar bem uma cena mas não representar bem outra cena e a mudança desse parâmetro gera muitas possibilidades de resultado. Então, utilizar métodos globais de ajuste de curva também não foi considerada uma alternativa viável principalmente pela enorme quantidade de dados a serem utilizados, demonstrando que a quantidade de dados também é um fator limitante para tentar algumas abordagens. Mesmo que o tempo não seja considerado importante nos testes, essa abordagem não leva a um resultado sólido pois a qualidade do resultado é muito influenciada por parâmetros livres.

4.5 Problema da Busca Cega

Diante de todas as dificuldades e de todas as análises realizadas neste capítulo chegou-se à conclusão que todos os problemas enfrentados para selecionar os pontos fazem parte de um problema ainda maior: **a realização de uma busca**

totalmente cega por um critério que defina a invariância dos pontos de saliência.

Essa busca é extremamente difícil porque podem ser testadas inúmeras possibilidades de técnicas de seleção de pontos e nenhuma delas pode resultar em uma resposta sólida sobre se é possível ou não selecionar pontos invariantes dentro de um mapa de saliência. Diferentemente dos detectores clássicos, não se sabe qual tipo de ponto é invariante, pois nos detectores clássicos os pontos são resultado de um filtro, e esse filtro procura ressaltar pontos que são apresentados com características bem conhecidas, então a seleção de pontos é simples e é conhecido a *priori* o que está sendo buscado e que o que está sendo buscado é invariante.

A busca por um critério de invariância não é explorada dentro da literatura da atenção visual, na verdade não há mapas de densidade construídos com imagens que sofreram diversas deformações. Então, as técnicas não têm uma referência para serem construídas com aspectos de invariância. Nesse sentido, há um questionamento ainda mais profundo a ser respondido, que é a respeito da invariância dos próprios mapas de densidade. Como não há trabalhos relatando esses aspectos nos mapas de densidade, que são a base para o desenvolvimento dos mapas de saliência, as técnicas computacionais nem chegam a se preocupar com esses aspectos. O único aspecto considerado pelos mapas de saliência é em apenas tentar se aproximar de forma grosseira dos mapas de densidade.

Como não se tem conhecimento a respeito do comportamento das próprias fixações humanas diante de imagens que passaram por transformações e por conta de todos os problemas mencionados que parecem indicar que os modelos computacionais ainda não estão preparados para extração local de características, o capítulo 5 faz uma investigação desses aspectos no sistema atencional biológico, por meio de uma pesquisa em bases de mapas de densidade existentes e de experimentos com seres humanos.

Capítulo 5

Investigação em Mapas de Densidade e Fixação

Como discutido no Capítulo 4, as investigações relacionadas a formas de extrair pontos de saliência que tenham características de invariância não são bem sucedidas em modelos computacionais de atenção, principalmente pelo fato dos modelos serem construídos sem levar esse aspecto em consideração. Computacionalmente, os mapas de saliência são idealizados apenas para limitar regiões de interesse, e para isso eles são construídos em baixa escala eliminando muita informação da imagem de entrada, o que torna a tarefa de extração ainda mais complexa.

O ideal é realizar a análise nos mapas de fixação humana, que representam de forma mais fiel os focos de atenção visual dos seres humanos. No entanto, como apresentado na Seção 2.3, não existem bases de mapas de fixação construídas para analisar aspectos de invariância das fixações com relação às transformações básicas que podem ocorrer em uma imagem. Como já mencionado anteriormente, tanto o sistema atencional humano como o computacional ainda não foram explorados em detalhes no aspecto de extração local de características.

Um dos principais empecilhos para a existência de uma base de mapas como essa é a quantidade de pessoas necessária para construir uma base simples e a quantidade de instâncias que cada imagem possui, resultando em um extenso trabalho de coleta de dados. No entanto, apesar dessas dificuldades e com o objetivo de extrair conclusões mais sólidas a respeito do potencial do sistema atencional para extração de características, foi construída uma base de mapas de fixação para analisar esse aspecto.

Na Seção 5.1 são explicados todos os passos e a metodologia utilizada para construir a base de mapas de fixação utilizada neste trabalho. Na Seção 5.2, é feita uma análise qualitativa e quantitativa dos dados obtidos, também como forma de validar a base de mapas criada.

5.1 Construção da Base de Mapas de Fixação

Com a construção de uma base própria de mapas de fixação espera-se alcançar melhor compreensão a respeito da presença da característica de invariância no processo pré-atentivo humano, e se os pontos provenientes das fixações ou dos mapas de densidade originados das fixações possuem invariância a algumas transformações na imagem.

Para organizar melhor a descrição das etapas de realização do experimento, da metodologia utilizada e da construção da base, a Seção 5.1.1 descreve apenas os materiais que foram utilizados na etapa experimental. A Seção 5.1.2 descreve a base de imagens utilizada para realizar o experimento, a Seção 5.1.3 descreve todos os passos do experimento e as decisões tomadas para definir a metodologia utilizada, a Seção 5.1.4 descreve o perfil dos voluntários, a Seção 5.1.5 apresenta o tratamento aplicado aos dados coletados e, por fim, a Seção 5.1.6 descreve os passos utilizados para construir os mapas de fixação e de densidade.

5.1.1 Equipamentos e *Softwares* Utilizados

Para a realização dos testes foi utilizado um equipamento de rastreamento ocular do tipo passivo, desenvolvido pela empresa THEEYETRIBE [1]. O equipamento foi disponibilizado pelo LAMID (Laboratório Multiusuário de Informática e Documentação) da Universidade Federal de Sergipe. É um equipamento de baixo custo, quando comparado a outros equipamentos mais robustos da área, e trabalha a uma frequência de 60 Hz. Ele é tido como um dos equipamentos de rastreamento ocular de baixo custo mais bem sucedidos do mercado de rastreamento e analisado como um equipamento viável para a realização de pesquisas científicas [87].

A Figura 5.1 ilustra o equipamento utilizado. Ele é composto internamente por um conjunto de LEDS infravermelho e de câmeras estéreo de alta resolução. As coordenadas da retina são encontradas a partir de informações do rosto e dos olhos e calculadas com relação à tela em que a pessoa está olhando. A informação do foco da retina é então representada por um ponto dado no sistema de coordenadas da tela em que as imagens são apresentadas, considerando que as imagens são apresentadas na mesma resolução da tela do computador utilizado.

Para a apresentação das imagens ao voluntário durante o teste, aquisição dos dados e calibração do sistema de rastreamento foi utilizado o *software open source* Ogama versão 5 [2]. O *software* foi utilizado em um computador DELL com o sistema operacional windows 10, processador Intel Core i5, 8 GB de memória RAM e 1 TB

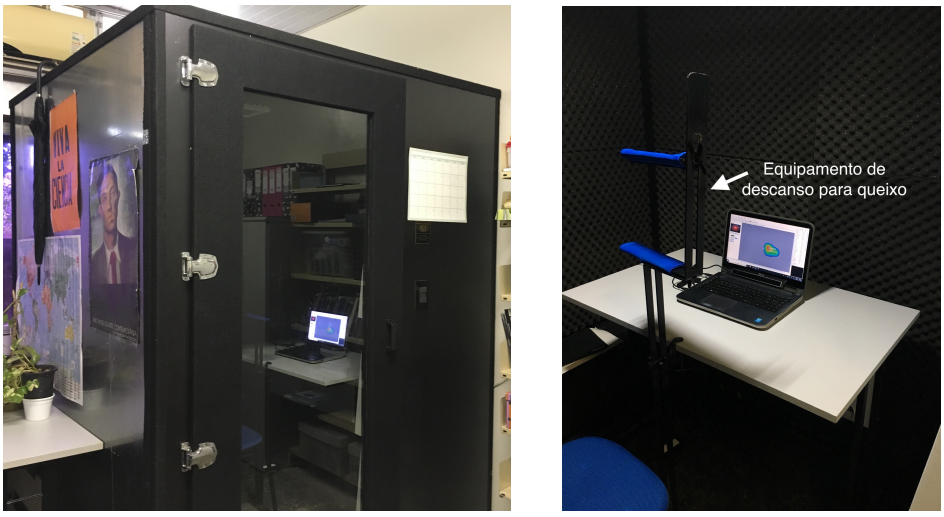
[1] Dados a respeito da empresa e do equipamento podem ser acessados em: `http://theeyetribe.com/dev.theeyetribe.com/dev.theeyetribe.com/general/index.html`

[2] Disponível gratuitamente em `http://www.ogama.net/node/3`

Figura 5.1: Equipamento de rastreamento utilizado.

de HD. Após realizados todos os testes, os dados foram exportados via arquivo de texto do Ogama para o MATLAB. A partir de então, as análises foram realizadas utilizando o *software* MATLAB versão 2015b.

Para a realização dos experimentos foi utilizada uma sala de isolamento acústico para atenuar a interferência de sons e ruídos do ambiente e também para manter a luminosidade sempre constante, já que esses fatores podem modificar o grau de concentração dos voluntários e por consequência interferir bastante na resposta de atenção visual. Além da sala de isolamento, também foi utilizado um suporte de queixo para fixar a posição da cabeça dos voluntários em uma mesma altura e impedir que ela se movimente ou que o voluntário fique cansado de se manter em uma mesma posição durante a visualização das imagens no teste. Toda essa infraestrutura também foi disponibilizada pelo LAMID, sendo que todos os testes foram realizados utilizando esse ambiente (Figura 5.2).



(a) (b)

Figura 5.2: Infraestrutura utilizada no experimento. (a) Cabine acústica utilizada no experimento para isolar os voluntários. (b) Equipamento para descanso de queixo utilizado no experimento.

5.1.2 Base de Imagens Utilizada

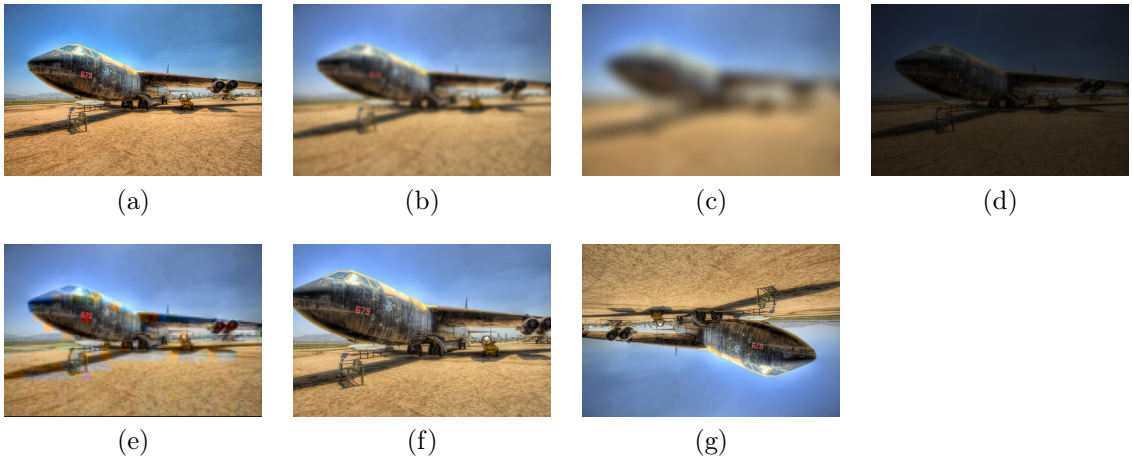
Para a realização do experimento foi necessário construir uma base de imagens própria, pois as bases existentes para esse propósito são muito reduzidas e não atendem às necessidades desse trabalho. A principal base de imagens da área, a *Oxford dataset* [3] apresenta cenas diferentes para cada transformação, ou seja, a cena que é apresentada para gerar a sequência de imagens com mudança de iluminação não é a mesma cena que é utilizada para gerar a sequência de imagens rotacionadas. Isso, de certa forma, prejudica a análise dos resultados, pois não é possível afirmar que um detector tem um comportamento superior ou inferior a uma determinada transformação, pois são cenas diferentes e o resultado é influenciado também pelo tipo de cena. Seja utilizando atenção visual ou detectores clássicos, a cena também influencia os resultados. Uma prova disso é que existem pesquisas destinadas somente a fazer comparações do desempenho dos detectores para resolver um determinado tipo de problema em um determinado tipo de cena [88].

As imagens utilizadas no experimento foram retiradas da base *COCO dataset* [4]. Essa base contém inúmeras imagens variadas, entre elas foram escolhidas 45 imagens para o experimento. As imagens escolhidas foram selecionadas para compor 2 grupos de teste: o **grupo I** e o **grupo II**. As imagens do grupo I possuem apenas um alvo atencional, sendo normalmente imagens de um objeto ou de um animal, esse grupo possui 21 imagens. As imagens do grupo II compõem imagens de ambientes externos ou de cenas com mais de um elemento atencional em destaque na cena, esse grupo possui 24 imagens.

Uma vez escolhidas as imagens foram realizadas transformações artificiais em cada uma delas. Foram feitas 6 transformações, fazendo com que cada imagem possua 6 instâncias. Foram feitas duas transformações de borramento, uma de compressão, uma de iluminação, uma de escala e uma de rotação.

Para realizar as transformações de borramento, as imagens foram convolvidas com um filtro gaussiano de desvio padrão 3,2 e outro de desvio padrão igual a 15. A compressão foi feita utilizando um algoritmo de compressão JPEG com fator de compressão igual a 24. As mudanças de iluminação foram produzidas escurecendo as imagens mantendo o brilho da imagem apenas com 25% do valor original. E por fim, a rotação foi feita também de modo artificial girando as imagens com um ângulo de 180°. Esse ângulo de rotação foi escolhido porque a disposição dos elementos da cena muda bastante e as pessoas têm mais dificuldade em reconhecer objetos na cena. A Figura 5.3 ilustra uma demonstração de uma imagem da base e suas 6 instâncias.

[3] Disponível publicamente em: `http://www.robots.ox.ac.uk/~vgg/research/affine/`
[4] Disponível publicamente em: `http://cocodataset.org/#download`



(a) (b) (c) (d)

(e) (f) (g)

Figura 5.3: Demonstração de uma imagem da base e suas 6 instâncias. (a) imagem original. (b) imagem com filtragem guassiana com um desvio padrão de 3,2. (c) imagem com filtragem gaussiana com um desvio padrão de 15. (d) imagem com apenas 25% da iluminação original. (e) imagem com compressão JPEG. (f) imagem com mudança de escala. (g) imagem rotationada.

5.1.3 Metodologia dos Testes

Inicialmente, para a realização do experimento foram considerados dois roteiros diferentes:

- **Roteiro I:** Uma sessão de 10 minutos por pessoa, com 90 imagens, sendo 15 delas a versão original das imagens e 75 um conjunto de 6 instâncias de cada uma das 15 imagens da base. Cada imagem permanece 3 segundos na tela, seguida por um tela preta com um ponto branco central que permanece 1 segundo. Isso totaliza 6 minutos de teste, sendo que os 4 minutos restantes seriam preenchidos com imagens sem instâncias, apenas para deixar o indivíduo mais tempo sem ver instâncias repetidas das imagens da base.

- **Roteiro II:** Uma sessão de aproximadamente 3 minutos por pessoa, com 45 imagens. Sendo as 45 imagens o grupo de imagens originais ou pertencentes a alguma transformação. Dessa forma, as pessoas não visualizam imagens repetidas e cada dia de realização de testes é realizado com imagens pertencentes a uma transformação diferente.

Os roteiros I e II foram testados com alguns voluntários antes de iniciar de fato os testes, para visualizar as vantagens e desvantagens de cada um.

O roteiro I apresenta como principal vantagem a necessidade de poucas pessoas, mas a sessão de teste é mais longa, e os voluntários consideraram cansativa, o que provoca mudança no nível de concetração das pessoas e afeta os resultados. Além disso, para uma sessão mais longa como essa, seria necessário recalibrar o sistema

pela menos duas vezes, uma no ínicio e outra na metade da sessão, para garantir a veracidade dos resultados. Um outro problema encontrado nessa abordagem é o efeito da memória das pessoas, pois apesar das imagens estarem embaralhas de forma aleatória e existirem imagens extras, as pessoas conseguiam identificar muito bem quando visualizavam imagens repetidas, e a partir daí, começavam a procurar algum tipo de padrão no teste, o que não é desejado, pois a influência da memória pode atrapalhar a análise de invariância, já que é procurado um padrão involuntário de invariância dos pontos no processo pré-atentivo.

Já o roteiro II apresenta como vantagens não apresentar imagens repetidas às pessoas, a sessão de testes é mais curta e não foi considerada cansativa pelos voluntários. No entanto, nessa abordagem é preciso de uma quantidade bem maior de voluntários diferentes, pois uma sessão de 3 minutos apresenta somente 45 imagens. Então, no mínimo seriam necessárias 273 pessoas para garantir uma análise de 6 instâncias de uma imagem, considerando que cada sessão seria testada com 39 pessoas no mínimo.

Apesar do roteiro I ser mais factível, foi utilizado nesse trabalho o roteiro II, porque ele não prejudica a veracidade dos resultados diante do que se deseja analisar. E, a partir de então, foram procurados muitos voluntários para realizar o teste.

No total, foram criadas 7 sessões do experimento, com 45 imagens cada. Cada imagem foi apresentada por 3 segundos na tela, seguida de uma tela preta com um ponto branco ou vermelho no centro da imagem, que ficou exposto por 1 segundo, com o objetivo de fazer as pessoas voltarem a focar no centro da tela, evitando a influência da resposta atencional da imagem anterior na próxima imagem que foi apresentada. O tempo de 3 segundos foi o tempo médio indicado na literatura para capturar a resposta atencional pré-atentiva do ser humano e por isso foi adotado nos experimentos [33]. Além disso, foi observado em testes extras antes da realização definitiva dos experimentos que a partir de 3 segundos a resposta das fixações de um ser humano começam a ficar repetitivas em torno dos mesmos pontos, por isso, não foi notada necessidade de aumentar o tempo.

Foi apresentada uma sessão com as imagens de forma original, duas sessões com as imagens borradas, uma sessão com as imagens comprimidas, uma com mudança de escala, uma com mudança de iluminação, e por fim, uma com as imagens rotacionadas a 180°. Sendo que, para cada sessão são necessários no mínimo 39 voluntários diferentes, respeitando a quantidade mínima de voluntários indicada na literatura [89]. Totalizando um quantidade mínima de 273 voluntários diferentes para criar a base.

Para iniciar a coleta de dados com cada voluntário, o computador foi posicionado a uma distância máxima de 60 cm e o primeiro passo realizado foi o de calibração, que visa coletar dados referentes ao mapeamento da fixação da retina nas coordenadas da

tela onde as imagens são apresentadas. Na etapa de calibração, um pequeno círculo apareceu e se moveu para 9 pontos específicos da tela (Figura 5.4). Então foi pedido que o voluntário seguisse o movimento do círculo com os olhos sem movimentar a cabeça. Feito esse procedimento, o *software* apresenta o grau de aceitação da calibração. Para garantir a qualidade dos resultados, as próximas etapas do teste só seguiram após a calibração de cada voluntário ser considerada no mínimo boa pelo sistema. Uma calibração boa significa que o erro da localização da fixação foi menor que 0,5° de ângulo visual.

Figura 5.4: Ilustração da etapa de calibração de 9 pontos.

O erro de calibração dos equipamentos de rastreamento ocular foi sempre dado numa medida de ângulo visual. Essa medida é referente ao tamanho do estímulo visual que é percebido pela retina, ou seja, o ângulo visual é o ângulo entre as linhas das extremidades do estímulo até a lente do olho, como ilustrado na Figura 5.5.

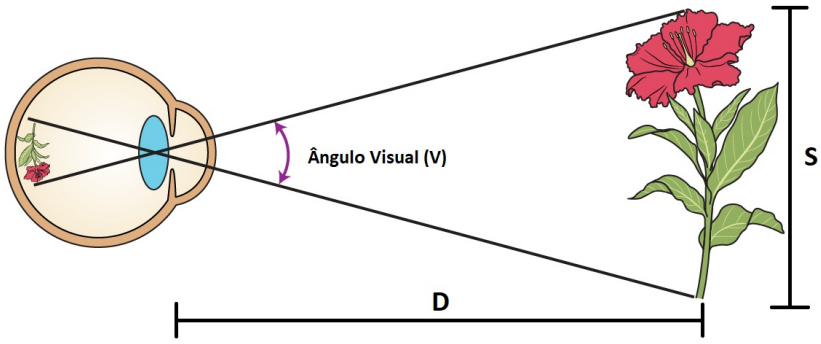


Figura 5.5: Representação da medida de ângulo visual.

Para encontrar o ângulo visual equivalente ao tamanho de um estímulo, foi utilizada uma relação simples de triângulos, dada por:

$$V = 2 \times tan^{-1}(\frac{S}{2D}) \tag{5.1}$$

Em que V é a medida em ângulo visual dada em radianos, S é o tamanho do estímulo dado em cm, e D é a distância do estímulo para o observador, também dada em cm.

Como pode ser visto por meio da equação 5.1, o valor do ângulo visual para um mesmo estímulo varia de acordo com a distância do observador. Isso mostra que o erro de calibração do ponto de fixação varia de acordo com a distância e com o tamanho da tela que é utilizada no experimento.

Para encontrar o erro de calibração das fixações que foram obtidas nesse experimento, a Equação 5.1 foi utilizada para encontrar quantos graus de ângulo visual são necessários para visualizar um *pixel* da tela. Para isso, é preciso saber a altura da tela que foi utilizada no experimento e a distância máxima em que o computador foi colocado do voluntário. A altura da tela utilizada foi de 18,6 cm e a distância máxima foi de 60 cm. Utilizando a relação expressa na equação 5.1, e convertendo o resultado para graus, o valor do ângulo visual foi de $\approx 17,6215°$. Esse valor foi o ângulo visual total utilizado para visualizar a tela de 18,6 cm. Para encontrar o valor equivalente em *pixels*, esse valor foi dividido pela resolução da tela utilizada, que foi de 768 *pixels*, resultando em um ângulo visual de $\approx 0,229°$ para visualizar um *pixel* da tela que foi utilizada.

Como o erro de calibração máximo foi de $0,5°$, fazendo uma regra de três simples, o equivalente a esse erro em *pixels* foi de 21,7916. Esse valor resulta em um raio de erro do ponto de fixação de ≈ 11 *pixels*.

Realizado o procedimento de calibração, os voluntários ficaram sozinhos na cabine de isolamento e foi pedido a eles que se concentrassem em observar todas as imagens durante todo o tempo em que elas foram expostas. Foi pedido que os voluntários visualizassem normalmente as imagens sem nenhum propósito específico. Isso foi feito para que a análise das pessoas sobre as cenas ocorresse da forma mais natural possível.

Ao final de cada sessão, foram obtidos os conjuntos de dados referentes à posição e duração das fixações de cada pessoa, para cada imagem do teste. Com esses dados, foi possível visualizar o comportamento do movimento ocular humano, como ilustrado pela Figura 5.6. Nessa figura, as fixações do voluntário estão no centro da área colorida, as linhas vermelhas indicam o movimento ocular. Já o tempo é representado pela intensidade da cor em volta do ponto de fixação. Quanto mais vermelhas e amarelas essas pequenas áreas, significa que a pessoa passou mais tempo com os olhos fixados naquele ponto.

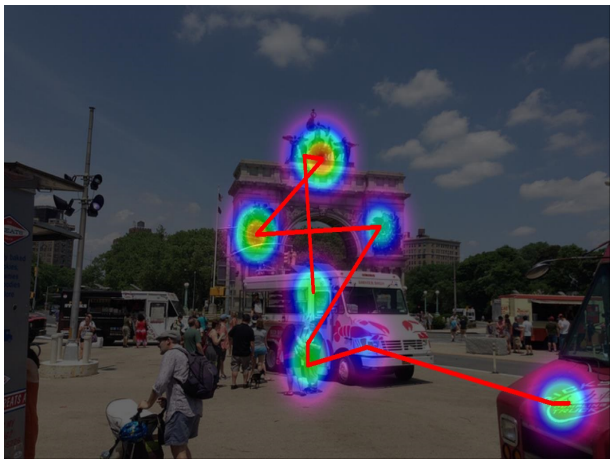


Figura 5.6: Representação visual das fixações e do movimento ocular de um voluntário, que foram obtidas durante o experimento.

5.1.4 Quantidade e Perfil dos Voluntários

O perfil e quantidade de voluntários utilizados em cada seção do experimento são apresentados na Tabela 5.1.

Sessão do Experimento (Transformação utilizada)	Quantidade de Pessoas			Faixa Etária
	Visão Normal	Problemas de Visão	Total	
Imagens Originais	43	5	48	(19-29)
Borramento 1	40	5	45	(18-50)
Borramento 2	40	3	43	(17-36)
Iluminação	37	7	44	(17-34)
Compressão	39	7	46	(17-50)
Escala	38	4	42	(17-40)
Rotação 180	38	7	45	(17-41)
TOTAL	275	38	313	

Tabela 5.1: Perfil dos voluntários.

No total, 313 pessoas diferentes participaram do teste, sendo 275 saudáveis e 38 afirmaram possuir algum incômodo na visão. Os dados das pessoas que reportaram algum problema de visão também foram coletados e foram analisados posteriormente na etapa de tratamento dos dados, que será descrita a seguir.

5.1.5 Tratamento dos Dados

Os dados brutos provenientes do rastreador ocular consistem em valores de tempo e posição de cada amostra para cada imagem de cada voluntário. Para evitar derivas de uma mesma fixação em regiões muito próximas da fixação original, o *software*

Ogama foi configurado para considerar fixações diferentes de um mesmo indivíduo quando as amostras estivessem a uma distância mínima de 10 *pixels*. Vale ressaltar que essa distância corresponde aproximadamente ao raio de erro da fixação adotado nesse trabalho. Essa filtragem impede que o Ogama registre como uma nova fixação um ponto que está dentro da margem de erro de um ponto já registrado anteriormente.

Além dessa filtragem que é feita pelos *softwares* de aquisição de dados, existem algumas abordagens para realizar o tratamento dos dados de fixação que consideram como derivas fixações ainda mais distantes que *10* pixels. Nessas abordagens são utilizados como critérios de descarte a distância, a velocidade da sacada e o tempo em cada ponto. Levando esses três critérios em conta, fixações de um mesmo voluntário ainda mais distantes que 10 *pixels* podem ser descartadas.

Esse critério mais elaborado de descarte de pontos, que não é feito pelo *software* de aquisição, mas sim em etapas posteriores de análise, é muito interessante para se construir mapas de saliência em que se deseja somente obter uma resposta grosseira da região em que as pessoas observam mais. No entanto, para um trabalho que tem como objetivo analisar extração local de pontos a partir da análise de fixações humanas não é um critério interessante, pois o fato de existirem fixações próximas a uma fixação inicialmente capturada, significa que existem informações muito relevantes nessas pequenas regiões da imagem que fazem com que a retina se fixe por mais tempo nos pequenos trechos das regiões próximas, e absorva ainda mais informação da vizinhança. Então, por esses motivos, foi considerado apenas o critério mais simples de descartar fixações redundantes, que é o critério de *10* pixels, pois ele está diretamente relacionado ao erro de calibração.

Além dos detalhes de filtragem das fixações, uma fixação só é considerada pelo equipamento quando ela passa um tempo mínimo de 100 ms, ou seja, quando são capturadas 6 amostras da mesma fixação, na mesma posição, em um tempo mínimo de 16,67 ms para cada amostra. Isso quando se trata do equipamento de 60 Hz, que é equipamento utilizando neste trabalho. Quando se trata do equipamento de 30 Hz, precisam ser capturadas 3 amostras, com um tempo mínimo de 33,33 ms cada, para que uma fixação seja considerada pelo equipamento. Coletados os dados, foi atribuído a cada ponto uma incerteza em sua posição, dando uma margem de erro de um raio de 11 *pixels*, cujo cálculo já foi discutido na Seção 5.1.3.

Um total de 313 voluntários distintos participaram do experimento, mas foi admitido que os voluntários pudessem realizar o teste novamente em outro dia, com o mesmo conjunto de imagens, só que com uma transformação diferente da já vista anteriormente. Então, além das 313 amostras conseguidas de cada pessoa, ocorreram 30 repetições referentes a 30 amostras de pessoas que repetiram o teste por uma segunda ou terceira vez. Esses dados provenientes de repetições foram analisados

Sessão do Experimento (Transformação utilizada)	Iniciantes		Experientes	Total de Pessoas
	Normais	Problema		
Imagens Originais	43	2	0	45
Borramento 1	40	3	3	46
Borramento 2	40	2	4	46
Iluminação	37	5	5	47
Compressão	39	5	4	48
Escala	38	2	4	44
Rotação 180	38	4	5	47
TOTAL	275	23	25	323

Tabela 5.2: Tabela com a quantidade de voluntários utilizada em cada sessão após o tratamento dos dados.

individualmente e foram considerados, pois não houve diferenças significativas das respostas das pessoas que repetiram o teste em outro dia com relação à resposta das pessoas que não repetiram o teste. Além disso, ao final do teste sempre era perguntado se as pessoas lembravam de muitas imagens. Alguns voluntários não lembravam de nenhuma, e outros chegaram a lembrar de no máximo 10 imagens. Como a resposta atencional não foi afetada, boa parte desses dados também foram considerados para a construção dos mapas de fixação.

Com relação às pessoas que informaram algum problema de visão, esses dados também foram tratados. Pessoas com problemas mais graves apresentaram um comportamento de velocidade das sacadas e de quantidade de pontos muito fora do padrão. Os dados dessas pessoas foram considerados *outliers* e descartados na versão final dos mapas de fixação. Ao final, após organizar os dados, a base foi construída utilizando a quantidade de candidatos indicada na Tabela 5.2. Nessa tabela, a quantidade total de voluntários que foi considerada por sessão para a construção definitiva da base é dividida em duas categorias principais: indivíduos iniciantes, ou seja, que realizaram o experimento pela primeira vez, e a de indivíduos experientes, que naquela sessão estavam repetindo o experimento por uma segunda ou terceira vez.

Como descrito na Tabela 5.2, a base foi construída com 323 amostras no total, considerando as amostras de pessoas que repetiram o experimento mais de uma vez em uma sessão diferente.

5.1.6 Construção dos Mapas de Atenção

Os mapas de fixação são transformados em mapas atencionais (ou mapas de densidade) a partir da modelagem dos pontos por meio de um conjunto de funções

gaussianas. A função gaussiana é representada matematicamente por:

$$G(x,y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{5.2}$$

Em que σ representa o desvio padrão da gaussiana, definindo a largura de abertura da função. Já o termo que multiplica a função exponencial define a altura do ponto máximo da função.

De acordo com a literatura, em cada ponto de fixação, de forma isolada, é atribuída uma gaussiana com desvio padrão $\sigma \approx 40$ *pixels*. Esse valor padrão sempre é utilizado na maioria dos trabalhos, e sua escolha é explicada pela medida de acuidade visual na região da fóvea, que é a região do olho que captura as informações em alta resolução. A acuidade visual é uma medida na área de oftamologia usada para avaliar a aptidão do olho em distinguir os detalhes espaciais. O que ocorre é que na região da fóvea a acuidade é muito alta e cai exponencialmente de acordo com o aumento do ângulo visual, como ilustrado na Figura 5.7. Quanto maior o ângulo visual, menor a acuidade da região da fóvea. Nesse experimento, para 1° de ângulo visual a acuidade é de $\approx 71\%$, e 1° de ângulo visual, de acordo com os cálculos na etapa de calibração é equivalente a um raio de ≈ 44 *pixels*.

Apesar dessa representação ser a mais conveniente para os mapas de densidade, pois ela representa muito bem a área de maior absorção de informação do olho humano, o desvio padrão utilizado para criar as gaussianas desse trabalho é definido como o raio do erro de calibração adotado, que foi de ≈ 11 *pixels*. Essa representação foi escolhida pois ela reflete melhor as características da superfície de saliência que são buscadas nesse trabalho.

Definido o desvio padrão, essas gaussianas são todas somadas para compor o mapa de densidade (Figura 5.8). Já a altura de cada gaussiana é proporcional à duração da fixação, ou seja, se um observador olhar para apenas um ponto na imagem durante o tempo em que a imagem é exposta, esse ponto irá representar o pico da gaussiana com altura igual a 1. Ao final desse processo, é criado um mapa bastante denso, que representa as regiões de maior foco atencional das pessoas, como ilustrado na Figura 5.8.

A representação da Figura 5.8 mostra um mapa muito denso, que é utilizado como mapa base na área de atenção visual, ou seja, todos os modelos computacionais tentam ter um aspecto parecido com esse mapa que representa misturas de gaussianas. E, de fato, essa representação está correta, mas esse modelo representa não só as fixações, mas também a parte mais significativa da visão periférica associada a uma fixação, então se trata também de uma representação grosseira de para "onde as pessoas estão olhando". Esse modelo esconde bastante as fixações,

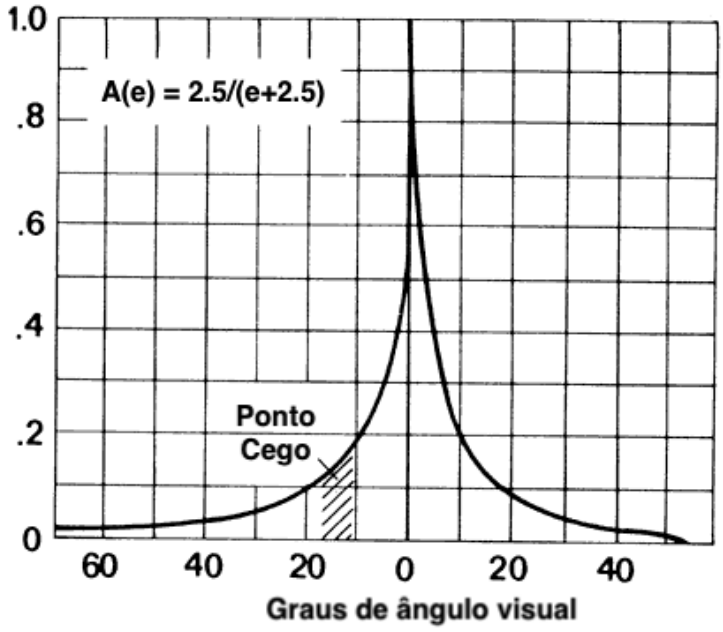


Figura 5.7: Representação da porcentagem de acuidade visual.[7].

elas acabam sendo sobrepostas por gaussianas de pontos mais próximos, dando um aspecto ao mapa de vários pontos igualmente importantes, formando grandes *platôs*. Por esse motivo, essa representação não é interessante quando se deseja analisar o conteúdo de saliência do ponto de vista de extração de características, pois os pontos do mapa acabam apresentando baixa distinção entre si e os máximos locais desaparecem, sendo que distinção é uma característica extremamente importante para extração de características.

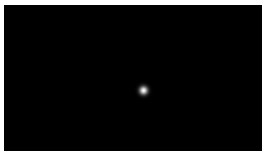
Como a área segue no sentido dos modelos computacionais replicarem de forma grosseira o modelo de atenção criado pelos mapas de fixação, que também é uma representação grosseira, isso traz a impressão aos pesquisadores que a utilização de mapas de saliência e seu conteúdo diretamente para extração de características é inviável, mas se trata apenas de uma questão de representação de modelo.

Por esses motivos, esse trabalho segue uma representação de modelo de atenção com o desvio padrão da gaussiana ajustado para o raio de erro máximo de calibração em torno do ponto de fixação. O raio de erro é de aproximadamente 11 *pixels*. Esse valor de raio é o valor adotado nesse trabalho para o σ da gaussiana, construindo um mapa de saliência menos denso no que diz respeito à quantidade de pontos conectados e com bastante distinção entre os elementos salientes, como ilustrado na Figura 5.9.

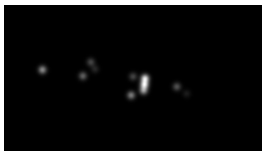
Os mapas construídos dessa forma serão utilizados nas seções seguintes deste trabalho para realizar comparações e avaliar a característica de invariância dos pontos de fixação.

Espera-se que com uma representação desse tipo, os pesquisadores da área se

(a)



(b)

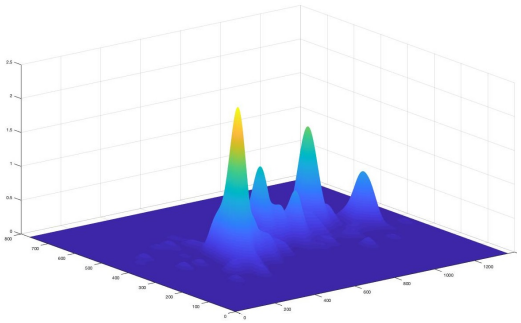


(c)

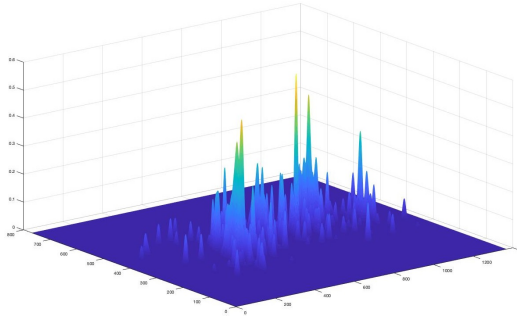
(d)

(e)

Figura 5.8: Construção dos mapas de atenção do modo tradicional com $\sigma = 40$. (a) imagem original. (b) Mapa de atenção composto por um ponto de fixação e uma gaussiana. (c) Mapa de atenção composto por 11 pontos de fixação e 11 gaussianas. (d) Mapa de atenção composto por 40 pontos de fixação e 40 gaussianas. (e) Mapa de atenção final composto por 345 pontos de fixação e 345 gaussianas.



(a)



(b)

Figura 5.9: Mapa de atenção com valores de σ diferentes. (a) Superfície tradicional construído a partir das fixacões, com $\sigma = 40$. (c) Superfície de densidade das fixações, com $\sigma = 11$.

atentem ao potencial do sistema atencional não somente para limitação de região de busca, mas também pelo fato de que existem muitos aspectos a serem explorados utilizando a informação de saliência do ponto de vista da extração de características locais. Essa nova representação traz uma nova visão para isso.

Definidos todos os detalhes de construção dos mapas atencionais e densidade de fixação, a próxima seção procura avaliar o potencial de invariância dos mapas, assim como realizar testes comparativos entre os resultados obtidos e os detectores clássicos, e uma comparação com os mapas de atenção que são considerados estado-da-arte da área.

5.2 Análise dos Mapas de Fixação

Essa seção é destinada a analisar os mapas obtidos durante o experimento. Os resultados apresentados visam responder aos seguintes questionamentos que foram acumulados no decorrer do trabalho devido às dificuldades encontradas:

1. O resultado dos mapas de fixação é comparável ao resultado obtido pelos detectores clássicos?

2. A atenção humana é guiada por um conjunto de elementos específico?

3. As pessoas olham para as mesmas regiões independente das transformações?

4. Como são caracterizadas as regiões da imagem que possuem maior concentração de fixações?

5. Quais as principais semelhanças e diferenças entre os pontos de fixação obtidos na imagem original e os pontos obtidos nas imagens que sofreram algum tipo de transformação?

6. Qualitativamente em qual(s) transformação(es) foram observadas maiores alterações de resposta das fixações?

7. Qual a porcentagem de pontos de fixação que se repetem entre as transformações, avaliando somente seu posicionamento (x,y) na imagem? Esses pontos se repetem nas mesmas regiões em todas as transformações?

8. Qual a porcentagem de pontos de fixação que se repetem entre as transformações, avaliando não só o posicionamento (x,y) e também o eixo z do mapa, que corresponde ao tempo da fixação. O tempo de fixação tem um papel relevante nos pontos que se repetem?

9. O resultado das análises obtidas varia entre o grupo de imagens? Ou seja, são obtidos resultados diferentes para as imagens do grupo I e para as imagens do grupo II?

10. É possível utilizar sistemas atencionais como extratores de características? Se sim, precisam ser realizadas modificações nos mapas? Quais modificações? Em quais modelos é mais factível realizá-las?

Todas essas perguntas são respondidas no decorrer de 3 Seções. Na Seção 5.2.1 são encontradas respostas para a pergunta 1. As perguntas de 2 a 6 são respondidas pelas análises da Seção 5.2.2, que apresenta análises qualitativas dos resultados obtidos. A Seção 5.2.3 responde às perguntas 7 e 8, mostrando uma avaliação quantitativa dos resultados obtidos. A pergunta 9 é respondida no decorrer tanto das análises qualitativas como quantitativas. Por fim, a pergunta 10 é respondida na Seção 5.2.4.

5.2.1 Regiões de Fixação x Detectores Clássicos

A referência de comparação para os detectores clássicos deveria ser as fixações humanas, pois não há como falar de extração de características visuais e reconhecimento de objetos e locais sem citar o mecanismo de atenção biológico. No entanto, não é o que ocorre na prática. Apesar de computacionalmente terem se desenvolvido duas áreas diferentes, uma para lidar com extração local de características e outra para lidar com atenção visual, e a junção das ferramentas das duas áreas não produzir resultados bons o suficiente, não significa que atenção e extração de características não tenham uma ligação forte biologicamente. Essa ligação só não tem sido bem explorada e bem representada computacionalmente.

Ao comparar os pontos encontrados pelos mapas de fixação com os pontos encontrados pelos detectores clássicos, pode-se avaliar que nem sempre os pontos coincidem nas mesmas regiões, e que o resultado é bastante relativo, ou seja, varia de imagem para imagem. Em cenas que existem muitas plantas e muita grama os detectores clássicos focam muito mais pontos nessas regiões, já que elas são mais ricas em gradiente local, enquanto que as fixações humanas não. Quando existe somente um objeto em destaque na cena, dependendo das características do objeto, os pontos podem ser mais coincidentes e às vezes totalmente diferentes.

Figura 5.10: Imagens com pontos resultantes de fixações e de alguns detectores clássicos. Em (a) (b) e (c) estão as imagens do grupo I com pontos detectados pelo Harris Corner Detector. Em (d) (e) e (f) estão as imagens do grupo I com pontos detectados pelo MSER. Em (g) (h) e (i) estão os pontos detectados pelo SURF. Em (j) (k) e (l) estão os pontos detectados pelas fixações humanas.

A Figura 5.10 ilustra alguns casos de teste em que as fixações estão centradas em

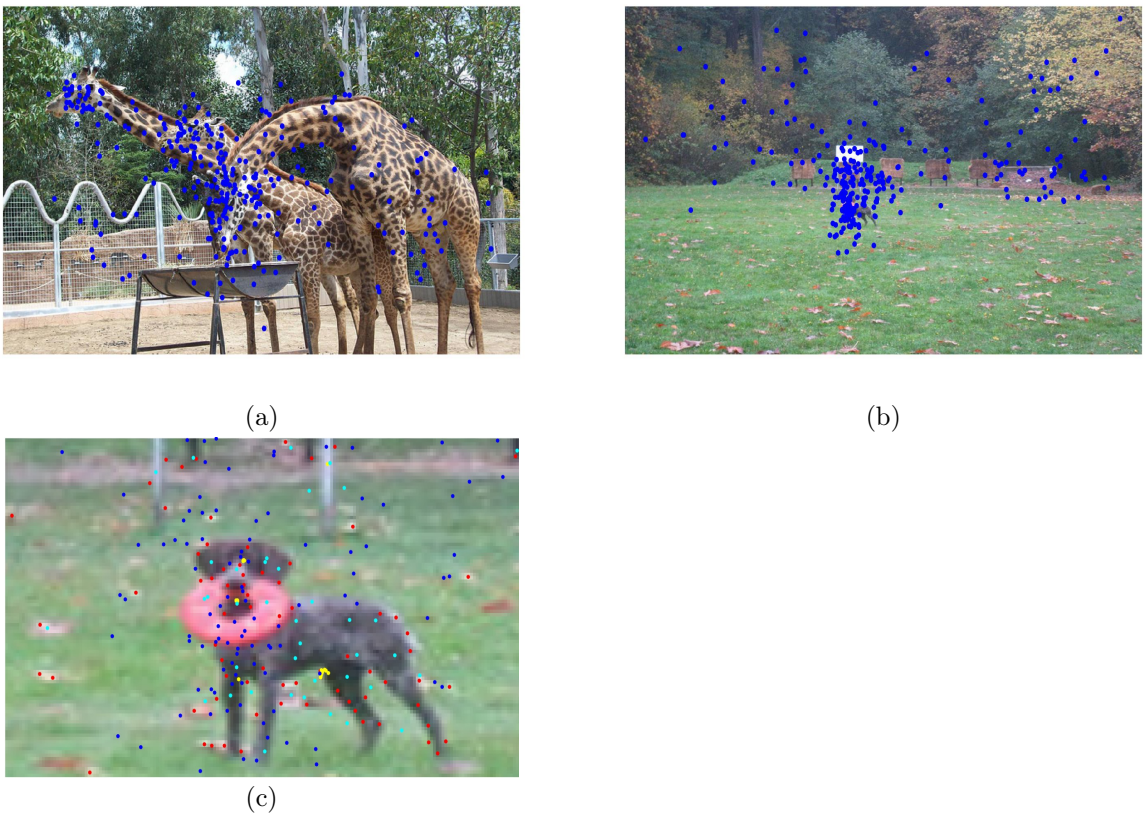
regiões diferentes das encontradas pelos detectores clássicos. Na Figura 5.10 em (j) é observado que enquanto as fixações estão em sua maior parte concentradas pela superfíce da torta, os detectores clássicos estão mais concentrados pelas bordas do garfo e da torta (Figura 5.10 (a), (d) e (j)). Há pouquíssimos pontos coincidindo com as regiões de saliência. Em um caso assim, se a região de saliência for utilizada como delimitadora de busca para os detectores clássicos, como é feito usualmente na prática, pouquíssimos pontos são encontrados dentro da região de saliência pelo detector, o que para a maiora das aplicações é ruim, pois um sistema com poucos pontos não possui robustez e está mais sucetível a erros. Na Figura 5.10 em (b), (e) e (h) ocorre um resultado semelhante, enquanto os pontos de fixação estão mais concentrados em torno de três regiões principais do urso, que são a face, o tronco e a etiqueta, os detectores estão muito mais concentrados nas orelhas e nos entornos dos braços e pernas. Apenas no rosto há alguns pontos coincidentes, principalmente com o SURF e com o Harris porque o sorriso do urso carrega muito gradiente local. Novamente, se a região de saliência for utilizada para delimitar a região de trabalho do detector, a região novamente será subutilizada e os resultados serão insatisfatórios.

(a) (b)

(c) (d)

(e) (f)

Figura 5.11: Imagens com pontos resultantes de alguns detectores clássicos. Em vermelho estão os pontos encontrados pelo Harris Corner Detector. Em amarelo estão os pontos encontrados pelo MSER. Por fim, em ciano os pontos encontrados pelo SURF. (a) Imagem do grupo II com pontos encontrados pelo Harris. (b) Imagem do grupo II com pontos encontrados pelo Harris. (c) Imagem do grupo II com pontos encontrados pelo MSER. (d) Imagens do grupo II com pontos encontrados pelo MSER. (e) Imagens do grupo II com pontos encontrados pelo SURF. (f) Imagens do grupo II com pontos encontrados pelo SURF.



(a) (b) (c)

Figura 5.12: Imagens com pontos resultantes das fixações humanas e de alguns detectores clássicos. Em azul estão os pontos resultantes das fixações humanas. (a) Imagem do grupo II com os pontos de fixação. (b) Imagem do grupo II com os pontos de fixação. (c) Região da imagem apresentada em (b) com pontos de fixação e pontos resultantes de detectores.

Se a imagem testada possuir elementos ricos em gradiente local pode ocorrer uma coincidência maior de pontos ou não, ainda depende muito do fato dos elementos que chamam a atenção humana serem os elementos ricos em gradiente local. Caso não sejam, as fixações não são coincidentes com os pontos encontrados pelos detectores. A Figura 5.10 ilustra um exemplo. Nessa Figura, em (l) as fixações estão mais concentradas na cabeça do cachorro, na parte interior do brinquedo e nas proximidades do carro. Os detectores por outro lado estão muito concentrados na grama, nas árvores e nos pontos que indicam o limite entre o chão e as árvores. No que diz repeito ao cachorro, os pontos ficam apenas em volta dele, mas raramente chegam próximos às localizações dos pontos de fixação, que estão concentrados na face.

Há casos em que o resultado é um pouco melhor, ou seja, em que há pontos detectados em regiões de alta concentração das fixações, mas para isso ocorrer, as áreas de fixação precisam ser muito ricas em gradiente local. Na Figura 5.11 são encontradas imagens em que isso ocorre. No entanto, as duas imagens são completamente pontuadas em todas as partes pelos detectores locais. Principalmente

na imagem das girafas, em (e), o SURF pontua praticamente toda a imagem, e as regiões da cabeça, em que há mais fixações, consequentemente também são pontuadas. Utilizando o Harris, os pontos ficam em torno do corpo da girafa e das plantas verdes de fundo. Já utilizando o MSER, os pontos ficam mais aglomerados pelo chão e pela cerca. Isso demonstra que a atenção humana não é somente guiada por gradiente. Se fosse, as fixações estariam espalhadas por todas as partes da imagem, mas isso não ocorre, há regiões bem específicas pontuadas pelas fixações humanas. Além disso, os detectores clássicos priorizam as partes de árvores e da grama. Há uma alta concentação de pontos nesses locais, e eles não são os mais relevantes para descrever o ambiente, na verdade eles são bastante redudantes e pouco significativos na descrição do ambiente. Além disso, como ilustrado na Figura 5.12 em (c), há uma concentração maior de pontos dentro da região das fixações onde está o cachorro, mas esses pontos raramente são coincidentes.

Na área de recuperação de imagens por conteúdo (CBIR) essa característica dos detectores clássicos em priozar elementos de fundo das cenas, como grama, árvores e o céu, é considerada crítica e até motivo de estudo de diversos pesquisadores, pois isso torna o sistema de busca lento e com muita informação que não descreve claramente a cena. O que é desejado para essa área, é o desenvolvimento de detectores e descritores compactos que representem bem o conteúdo mais significativo da cena, para agilizar a busca. O desenvolvimento de um detector inspirado nas fixações humanas poderia contribuir bastante para os problemas dessa área nesse sentido.

Como descrito nessa seção, é notável que os detectores busquem características distintas das características buscadas pelo sistema atencional, e que ainda há muitas coisas relevantes a serem exploradas dentro da extração local de características. Além disso, a área de extração de características poderia utilizar os mapas de fixação como referência para novas descobertas.

5.2.2 Análise Qualitativa dos Mapas

De um modo geral, o sistema atencional humano é atraído por um conjunto de características de alto nível. Além disso, é observado que existem algumas prioridades para as fixações humanas.

O seres humanos são atraídos principalmente pelo conjunto de características que compõem ou parecem compor algo semelhante a uma face. Se existem seres humanos na cena ou animais, a cabeça é o alvo prioritário das fixações, pois as fixações demoram mais tempo nela. Se existem seres inanimados na cena, como objetos ou veículos, a tendência está sempre em olhar para as partes do objeto que se assemelham também com uma cabeça. Após observar a cabeça, existe uma tendência em estimar o comprimento do corpo, passando as fixações da cabeça para

o tronco e posteriormente na direção dos membros inferiores, quando se trata de pessoas ou animais, ou para a parte traseira ou da base, quando se trata de veículos ou outros objetos.

Para ilustrar essa observação, a Figura 5.13 mostra algumas cenas com a representação das áreas de fixação selecionadas por seis voluntários diferentes, ou seja, cada cena traz as fixações de apenas um voluntário. As áreas coloridas representam as regiões de maior foco atencional da pessoa, sendo que no centro da região está a localização aproximada da fixação. Quanto mais amarela e vermelha a área, significa que o voluntário fixou mais tempo os olhos na pequena região colorida, já as linhas vermelhas indicam o caminho percorrido pelos olhos entre as fixações. Essa representação foi feita apenas com as fixações de um voluntário por imagem com o objetivo de não poluir a ilustração das cenas e facilitar a explicação do que foi observado, já que muitos voluntários fixaram a atenção em regiões semelhantes e apresentaram caminhos de fixação também semelhantes.

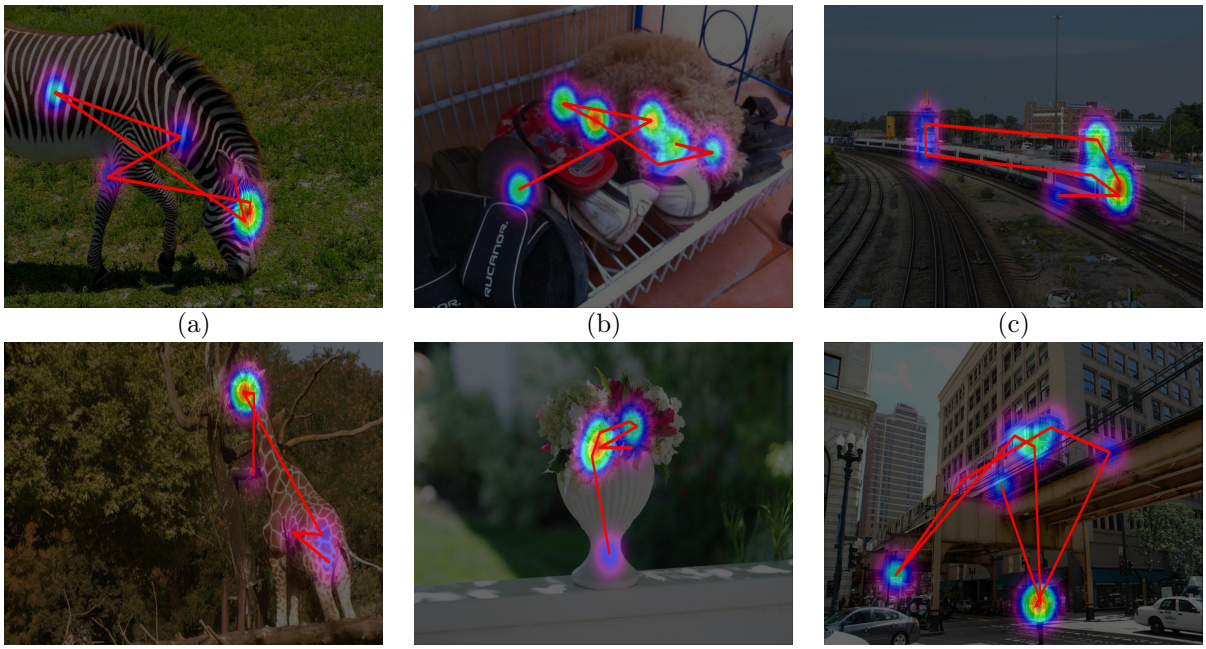


Figura 5.13: Ilustração de focos atencionais de um voluntário por imagem utilizando as imagens originais. (a) Cena de uma zebra pertencente ao grupo I de imagens de teste. (b) Cena de elementos em uma cesta pertecente ao grupo II de imagens de teste. (c) Cena de um trem pertencente ao grupo II de imagens de teste. (d) Cena de uma girafa pertecente ao grupo I de imagens de teste. (e) Cena de um jarro de flores pertencente ao grupo I de imagens de teste. (f) Cena de um metrô pertencente ao grupo II de imagens de teste.

Nas imagens (a) e (d) da Figura 5.13 é perceptível que os voluntários passam mais tempo analisando os detalhes da cabeça que do corpo da zebra e da girafa, mas que também fazem uma varredura mais rápida do corpo aparentemente no sentido

de estimar o comprimento. Essa característica também é observada nas imagens que estão em (c) e (f) (Figura 5.13), com a diferença que o foco atencional está voltado em analisar a parte da frente do trem e do metrô, como se fossem faces, e principalmente na imagem do trem, em (c), é perceptível que após passar mais tempo visualizando a parte da frente a varredura visual se volta para a parte traseira do veículo retornando para a parte frontal do trem novamente. Já na imagem do metrô a fixação se inicia dando foco principal a uma pessoa na cena, principalmente em sua face. Em seguida, a fixação segue para a parte frontal do metrô e há uma varredura maior destacando três focos atencionais nessa região. A partir da parte frontal aparece uma nova fixação em direção à parte traseira e posteriormente o olhar do voluntário vai para uma variedade de carros e pessoas embaixo da ponte e volta para a parte frontal do metrô. Por fim, a varredura da cena é finalizada no ponto de partida, ou seja, na face do indivíduo na rua.

Nas imagens (b) e (e) também da Figura 5.13, o padrão de visualização não é muito diferente. Na imagem em (b), apesar de existir uma sandália vermelha se destacando bastante no ambiente, é dada uma certa prioridade a investigar primeiro as características do cão, do local onde sua face está escondida e do seu corpo. Em (e) os primeiros focos atencionais estão voltados a analisar a parte frontal do objeto e, em seguida, segue uma fixação na direção da base do jarro.

Outro fator interessante é que o foco nessas regiões é aparentemente independente do voluntário e independente da transformação ocorrida na imagem. Para demonstrar isso visualmente, as Figuras 5.16 e 5.17 ilustram o conjunto de todas as fixações das imagens que foram apresentadas na Figura 5.13. Essas fixações foram extraídas de todos os voluntários que participaram da sessão de testes com as imagens em sua forma original e transformadas. Em todas as imagens da Figura 5.16 e em todas as imagens da Figura 5.17 a maior parte das fixações está organizada em torno das mesmas regiões delimitadas pelo foco atencional de apenas um voluntário, indicado nas imagens da Figura 5.13, e isso não ocorre somente com a imagem em sua forma original, mas também em todas as outras instâncias das imagens originais, demonstrando que o sistema atencional é guiado por características de alto nível independentemente de algumas deformações locais que ocorrem na imagem.

Nas transformações de borramento, muitos detalhes de alta frequência são perdidos na imagem e mesmo assim as fixações continuam em torno das regiões com características de faces e seguem explorando o corpo dos elementos. Vale ressaltar que a consistência das fixações em torno dessas regiões nos dois níveis de borramento testados são qualitativamente boas, pois as fixações estão em torno das mesmas regiões seguindo o mesmo padrão. No entanto, é perceptível uma leve diferença entre as fixações da imagem original e as fixações das imagens borradas. Nota-se que há uma redução na quantidade de fixações das imagens borradas, o que é compreensí-

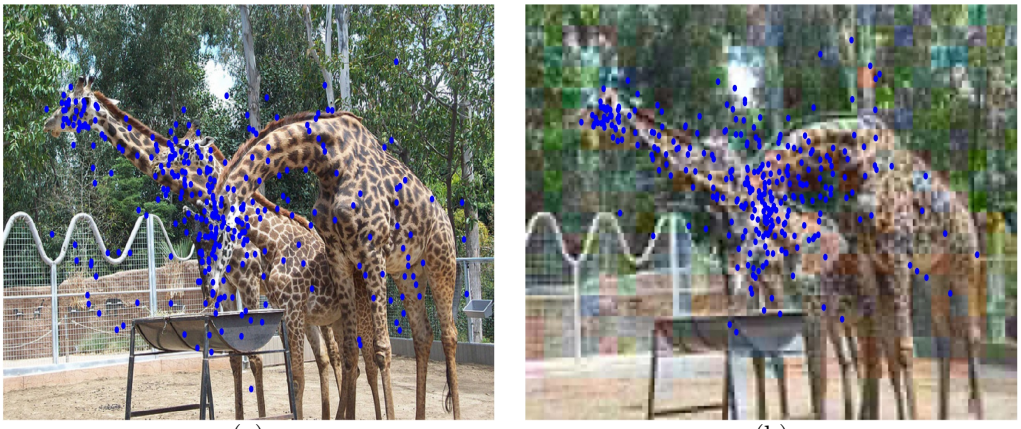
vel, pois as pessoas tem menos o que olhar nas imagens e ficam também confusas tentando decifrar do que trata a cena principalmente quando o nível de borramento aumenta. Além disso, as pessoas não esperavam olhar para cenas borradas. Apesar de todos esses fatores, é perceptível que há muitos pontos nas mesmas regiões tanto em imagens do grupo I como em imagens do grupo II.

Outra observação importante ocorre quando a imagem é girada 180°, esse ângulo foi escolhido justamente porque as pessoas sentem maior dificuldade em interpretar a cena quando está girada a esse ângulo. Quando as imagens giradas foram utilizadas com os voluntários nos testes, eles saiam da sala afirmando sentir mais dificuldade que o comum para interpretar as imagens, pois não era algo natural para as pessoas e elas tinham que se concentrar mais que o comum para entender a cena. Então, era esperado que talvez a rotação fosse uma característica aprendida pelo ser humano, e que as fixações fossem muito diferentes, já que as pessoas precisavam de mais concentração, era esperado que observassem a cena de outra forma, mas isso não ocorreu na prática. O que ocorreu foi que as fixações seguiram o mesmo padrão de busca, e como foram testes realizados com pessoas diferentes, nota-se que a rotação não é uma característica aprendida, mas que há invariância do sistema atencional com relacão a rotação, pois acredita-se que sob rotações menores os indivíduos vão sentir ainda menos dificuldade para entender a cena, e será um processo ainda mais natural.

Os resultados trazidos pela rotação são bastante interessantes porque mesmo nas imagens do trem e do metrô, que são mais complexas e que dão mais possibilidade das pessoas olharem para outras regiões, as prioridades dos focos atencionais foram as mesmas ou até melhores que os resultados obtidos pelas outras instâncias das imagens.

Nas tranformações de escala, compressão e iluminação também foram observados bons resultados. As imagens mais escuras tornam alguns elementos que antes apresentavam um contraste maior com relação ao ambiente menos contrastantes, como os prédios na imagem do trem. Apesar disso, os resultados continuaram seguindo o mesmo padrão. Nas transformações de compressão o principal desafio foi que regiões da imagem ficavam um pouco desfiguradas e mesmo assim as fixações se mantinham. Um exemplo disso pode ser visto na Figura 5.14 em (b), em que apesar da dificuldade para identificar claramente os detalhes das girafas, as fixações estão em torno das mesmas regiões.

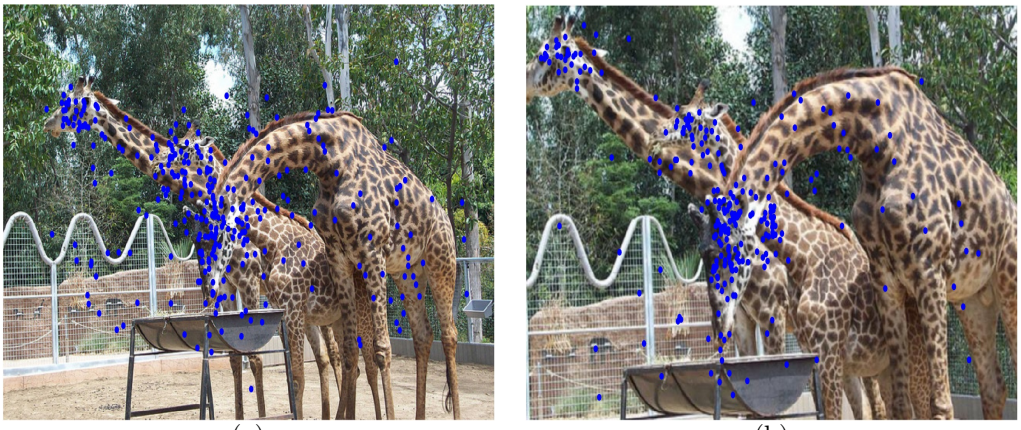
Nas transformações de escala, apesar da mudança ter sido um pouco sutil, era esperado que ocorresse um aumento de fixações, e que mais detalhes das regiões de alta atenção fossem capturados, já que alguns elementos desapareceram da cena, e as pessoas teriam supostamente mais tempo para varrer os focos atencionais mais cuidadosamente. Além disso, os detalhes de alta frequência das regiões de atenção



(a) (b)

Figura 5.14: Imagens com seus respectivos pontos de fixação. (a) Imagem original. (b) Imagem comprimida.

ficam mais visíveis para os indivíduos. No entanto, isso não ocorreu na proporção imaginada, ou seja, não houve um aumento significativo de fixações em uma determinada região, o que demonstra que o sistema atencional humano não é guiado prioritariamente por detalhes de alta frequência da imagem. A Figura 5.15 em (b) ilustra uma imagem com mudança de escala e é perceptível que, apesar de alguns elementos não estarem mais presentes na cena, e existirem mais detalhes que podem ser percebidos pelos olhos, não há um aumento significativo das fixações quando comparadas com os resultados obtidos na imagem original (Figura 5.15 em (a)).



(a) (b)

Figura 5.15: Imagens com seus respectivos pontos de fixação. (a) Imagem original. (b) Imagem com mudança de escala.

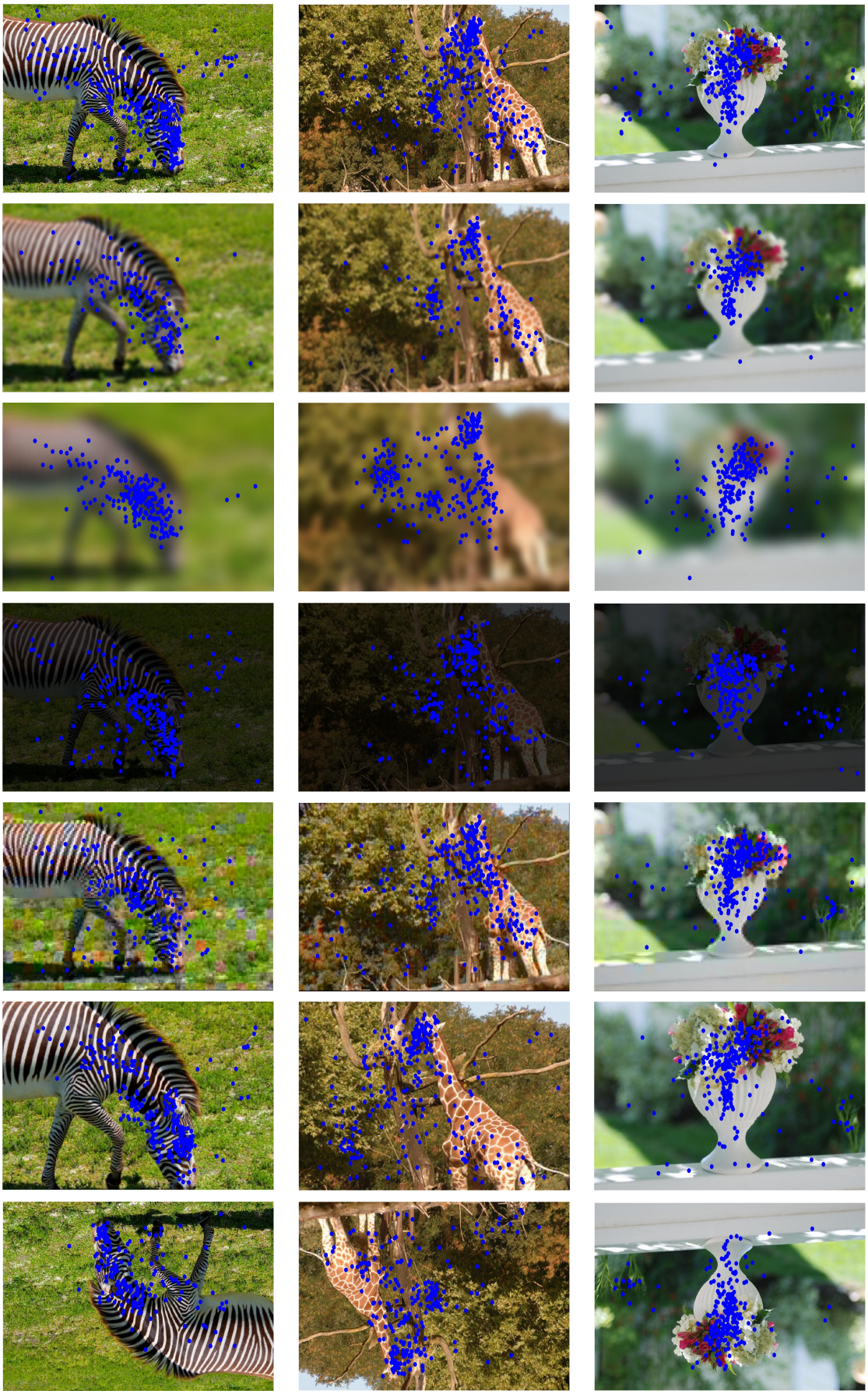


Figura 5.16: Imagens com seus respectivos pontos de fixação. Na primeira linha se encontram as imagens na versão original, seguidas pelas seguintes transformações: borramento, iluminação, compressão, escala e rotação.

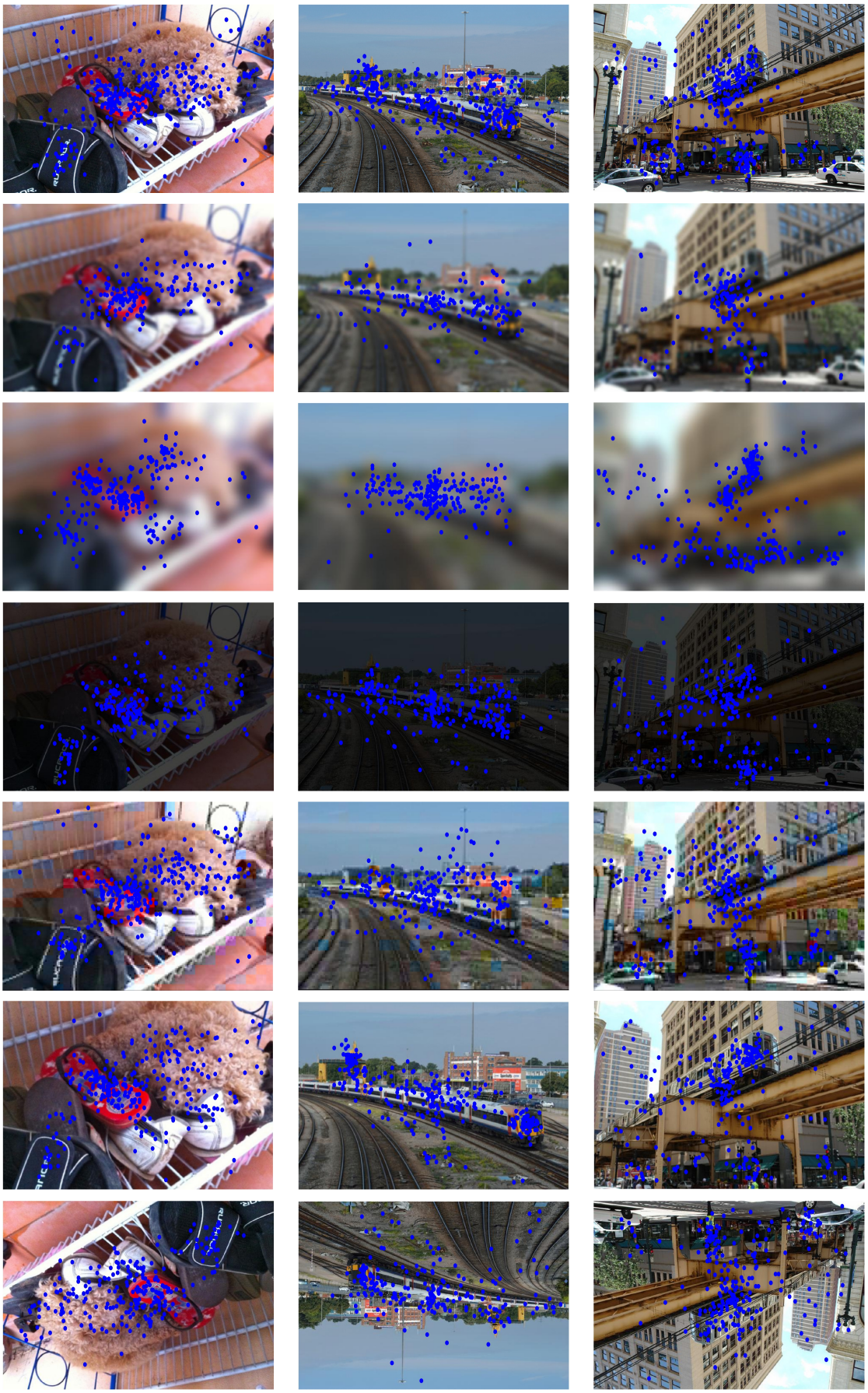


Figura 5.17: Imagens com seus respectivos pontos de fixação. Na primeira linha se encontram as imagens na versão original, seguidas pelas seguintes transformações: borramento, iluminação, compressão, escala e rotação.

Vale ressaltar também que todas as observações citadas até o momento ocorrem tanto no grupo I quanto no grupo II das imagens. Além disso, esse comportamento demonstra evidências de que a atenção é um mecanismo de sobrevivência e que pode ser institivo o fato de procurar elementos semelhantes a faces e tentar estimar o comprimento desses elementos para verificar rapidamente se existe algum perigo iminente ao estímulo que foi apresentado visualmente, ou se é algo inofensivo.

Figura 5.18: Imagens com seus respectivos pontos de fixação. Na primeira linha se encontram as imagens na versão original, seguidas pelas seguintes transformações: compressão e rotação.

Os seres humanos também são atraídos por palavras e caracteres, inclusive alguns voluntários chegaram a comentar que tentaram ler as palavras que apareciam eventualmente em algumas imagens. Sempre que a imagem, mesmo em condição ruim, davam indícios que existiam palavras, alguns voluntários afirmaram naturalmente tentar decifrar o que estava escrito. Talvez esse seja um comportamento cultural que influencie o sistema atencional a se fixar em palavras na cena independente da condição da imagem. Esse comportamento é observado na Figura 5.18. Nas imagens do ônibus as pessoas focaram em olhar para a maior parte dos textos escritos, estando eles no ônibus, que é o elemento de maior destaque, ou não, como os textos que estão na placa em azul e atrás do ônibus. Na imagem comprimida do ônibus não é possível entender corretamente todos os caracteres escritos e mesmo assim os focos atencionais continuam voltados para os mesmos elementos. Além disso, tanto na imagem do ônibus como na da placa, quando rotacionadas, é interessante observar que as fixações ficam em sua maioria em torno das mesmas letras.

Outra prioridade dos seres humanos é a de fixar bastante atenção em todas

as partes da cena que possuem comida. Quando existem muitas faces e comida, aparentemente a prioridade está em fixar atenção na comida e posteriomente varrer as faces ou elementos parecidos com faces presentes na cena. Quando existem muitas comidas dispostas na cena, praticamente todas as regiões que possuem comida são varridas pelas sacadas humanas. Um comportamento que também é independente da situação da imagem, como ilustrado pela Figura 5.19.

Figura 5.19: Imagens com seus respectivos pontos de fixação. Na primeira linha se encontram as imagens na versão original, seguida pelas imagens rotacionadas na segunda linha.

Como relatado nessa seção, não foram observados desvios de focos atencionais quando a imagem sofre alguma transformação. Apesar de não terem sido realizados testes com uma ampla quantidade de imagens, o fato dos seres humanos se fixarem em elementos semelhantes a faces, sejam elementos vivos ou não, cobre uma grande variedade de cenas, tanto internas quanto externas. Além disso, para as transformações que foram testadas nesse trabalho, não ocorreram mudanças significativas de foco nem do comportamento das fixações, apenas mudanças sutis de comportamento no conjunto de fixações referentes a imagens borradas. Isso é perceptível visualmente (Figuras 5.16 e 5.17), mas são mudanças pequenas. Isso sugere que qualitativamente o borramento seja a transformação em que as fixações sofram mais com relação à sua consistência, e que isso depende do grau de borramento.

Com relação às regiões em que as fixações se aglomeram em maior quantidade, elas podem ser regiões mais lisas, que fazem parte do corpo dos objetos ou pessoas, ou regiões com um pouco mais de gradiente e de textura, mas muitas vezes esse gradiente não é considerado relevante para os sistemas computacionais de visão que

buscam por gradiente, como foi discutido na seção 5.2.1. Então, há uma variedade bem maior de possibilidades que não pode ser representada pela busca de uma característica única dentro das regiões de foco atencional da imagem. Quando se trata de faces, as fixações costumam aparecer mais em torno da boca e olhos, ou algo semelhante a eles quando se trata de objetos. Quando se trata de palavras, as fixações ocorrem em torno dos cantos das letras em maior quantidade. Quando se trata de plantas as fixações ocorrem em maior quantidade em regiões de maior contraste de cor, quando esse constraste existe.

Também foi observado que o sistema atencional humano não é drasticamente afetado por transformações locais nas imagens, na verdade ele parece ser guiado por um conjunto de características de alto nível que são pouco afetadas por alterações médias de luminosidade, de inversão do posicionamento dos elementos na cena, pela degradação parcial dos elementos da cena ou pela ausência de detalhes de alta frequência. Isso é refletido principalmente ao olhar para os resultados da transformação de rotação, em que as imagens são invertidas e não se nota nenhuma diferença significativa na aglomeração de fixações nas regiões da imagem quando comparadas às fixações da imagem original. Esse mesmo comportamento também se reflete nos resultados das outras transformações, mostrando uma consistência relevante das fixações.

Do ponto de vista qualitativo, e de acordo com os casos de teste que foram utilizados nesse trabalho, sugere-se que exista potencial para sistemas atencionais como extratores locais de características, e que as fixações são ótimas candidatas a pontos-chave. Para avaliar melhor outros resultados qualitativos obtidos, as imagens com as fixações e os mapas de densidade estão disponíveis para acesso nesse trabalho [5]. Além disso, para validar quantitativamente os resultados e verificar o quanto as fixações são afetadas pelas transformações, a Seção 5.2.3 apresenta os resultados quantitativos obtidos.

5.2.3 Análise Quantitativa

Como discutido na Seção 5.2.2, as fixações tendem a ir para as mesmas regiões da imagem independente das transformações em que foram testadas. No entanto, qualitativamente são observadas diferenças sutis entre as fixações provenientes das imagens originais e provenientes das imagens que sofreram algum tipo de transformação. Essa seção tem como objetivo avaliar o quanto essas diferenças afetam quantitativamente a consistência das fixações.

O primeiro fator de diferença entre as fixações provenientes das imagens originais e as fixações provenientes das imagens transformadas é a quantidade de fixações

[5] Acesso às imagens em: `https://mega.nz/#F!5B9SzCCZ!byqrHYW4XqRL83MVD-IB5Q`

obtidas. De um modo geral, ocorre uma redução na quantidade de fixações nas imagens transformadas, sejam transformações mais sutis, como a de mudança de escala, ou transformações mais complexas como borramento e rotação 180°. A Figura 5.20 apresenta o impacto da queda da quantidade de fixações em cada transformação para cada imagem que foi utilizada no teste.

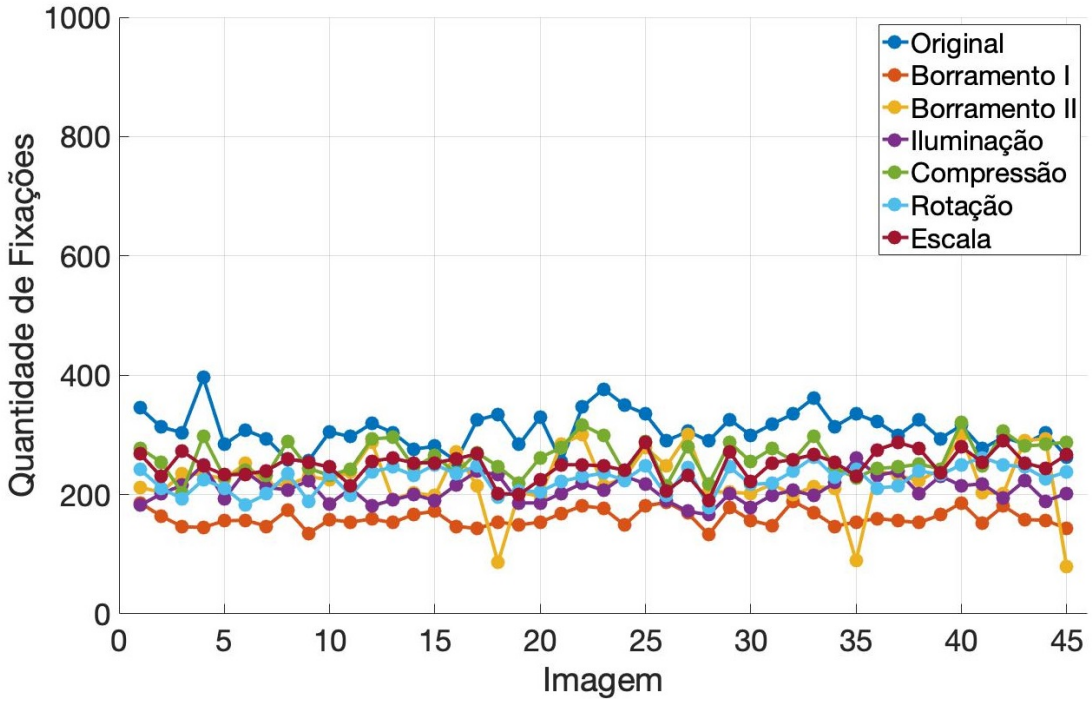


Figura 5.20: Quantidade de Fixações para cada imagem da base.

Como apresentado no gráfico da Figura 5.20, a quantidade de fixações para as imagens originais varia em torno de 300 a 400 pontos por imagem. Já para as imagens transformadas, ocorre uma redução maior que 100 pontos a depender da imagem e da transformação. É compreensível que exista um impacto na quantidade de pontos em algumas transformações. Nas transformações de borramento e compressão, muita informação é eliminada da imagem, há menos elementos nítidos para analisar, então é mais difícil entender a imagem e as pessoas não esperam ver as imagens nessas condições, pois não é natural para elas. Esses fatores deixam as pessoas mais confusas, o que era desejado no experimento. Nas transformações de rotação e iluminação também é compreensível que existam modificações com relação à quantidade de fixações, pois não é natural para o ser humano avaliar uma imagem invertida. Já no caso da transformação de iluminação, as pessoas podem não se sentir tão desconfortáveis em analisar as imagens, mas há uma redução significativa de contraste entre os elementos. Já com relação à mudança de escala, era esperado o inverso, pois os detalhes dos elementos nessas imagens são mais perceptíveis. No entanto, também ocorreu uma redução significativa em algumas imagens.

Apesar da redução da quantidade de fixações em todas as transformações, os pontos de fixação encontrados estão em torno das mesmas regiões, dando bons indícios de invariância do sistema atencional humano, como foi discutido na seção 5.2.2. Apesar disso, é preciso verificar o quanto essas pequenas diferenças afetam a repetibilidade dos pontos de fixação, com o objetivo de concluir de forma quantitativa a respeito da invariância dos pontos, e se eles realmente são bons candidatos a pontos-chave.

A repetibilidade é um aspecto muito importante em detectores de características locais. Ela se trata da porcentagem de pontos que se repetem nas mesmas posições, quando são comparados pontos encontrados em uma imagem que não sofreu transformações com pontos de uma imagem que sofreu algum tipo de transformação. Dessa forma, a repetibilidade é o que define a qualidade de um detector na área de extração de características.

Para calcular a repetibilidade, os pontos da imagem original são projetados na imagem transformada (Figura 5.21). Para fazer isso é preciso saber qual foi a transformação geométrica sofrida pela imagem, ou seja, é preciso ter acesso à matriz de homografia que faz esse mapeamento corretamente. Quando se tratam de transformações que não afetam o posicionamento dos pontos, como borramento e iluminação, por exemplo, não há necessidade da matriz de homografia, pois o mapeamento é direto, já que a posição dos pontos não foi alterada.

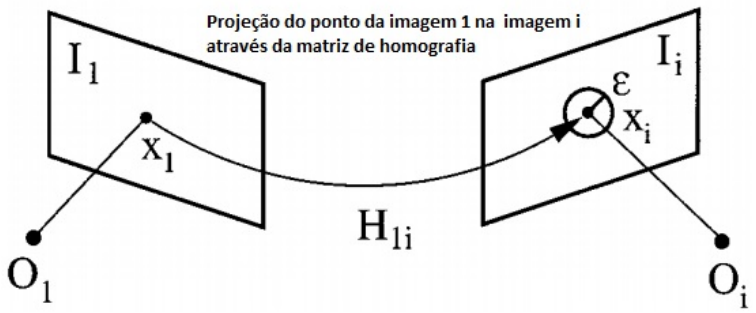


Figura 5.21: Ilustração da etapa de projeção de pontos entre duas imagens utilizando a matriz de homografia [8].

Após projetar um ponto da imagem I_1 na imagem I_i é checado se esse ponto projetado está próximo de algum ponto da imagem I_i considerando uma área de erro, que reflete a incerteza da detecção. Se o ponto projetado estiver próximo de algum ponto nessa área isso é considerado uma correspondência e é contabilizada na métrica de repetibilidade. Esse processo é feito com todos os pontos da imagem I_1 e ao final do processo se obtém o total de correspondências que ocorreram. A partir daí, é criada uma relação entre o total de correspondências encontradas e o total de pontos que foram detectados, resultando em um percentual de acerto do detector. Matematicamente, essa taxa é expressa da seguinte forma:

$$r_i(\epsilon) = \frac{T}{min(n_1, n_i)} \tag{5.3}$$

Em que $r_i(\epsilon)$ é a repetibilidade dos pontos considerando uma janela de erro ϵ, T é o total de correspondências encontradas a partir da projeção dos pontos da imagem I_1 na imagem I_i e $min(n_1, n_i)$ é o menor valor entre a quantidade de pontos detectados na imagem I_i e a quantidade de pontos detectados na imagem I_1. Essa métrica considera que o total de correspondências deve ser dividido pela menor quantidade de pontos, pois em transformações de escala e perspectiva, muitos pontos que existem na imagem I_1 não existem na imagem I_i e seria uma comparação injusta nessas condições, por isso serão propostas algumas modificações nessa métrica.

A primeira modificação realizada nessa métrica foi para tratar de casos de redundância de pontos de fixações. Em um mapa de fixação às vezes existem aglomerações de pontos em uma mesma micro-região da imagem. Claro que essas aglomerações não são tão constantes como em um mapa de densidade convencional apresentado na literatura, pois há distinção e mais separação entre os pontos, mas também há situações em que os pontos estão realmente conectados, representando praticamente a mesma coisa. A Figura 5.22 ilustra um trecho de pontos extraído de um mapa de fixação para mostrar casos em que os pontos parecem ser redundantes e casos em que isso não ocorre.

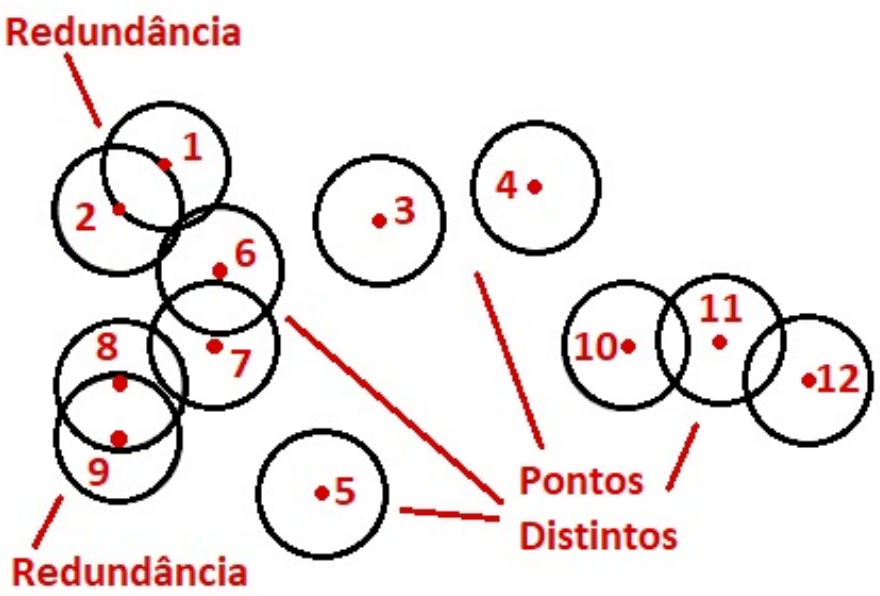


Figura 5.22: Trecho de pontos extraído de um mapa de fixação.

A Figura 5.22 ilustra 12 pontos de fixação extraídos de um mapa que foi obtido durante os testes. Nessa figura, a posição do ponto obtida pelo equipamento é re-

presentada pelos pontos vermelhos dentro da região circular, e a região circular é representada pelo raio de erro máximo obtido no experimento para cada ponto, que também é a região da gaussiana do mapa de densidade considerado nesse trabalho. Como ilustrado nessa figura, as fixações 1 e 2 são praticamente redundantes, pois o centro da região circular da fixacão 1 está exatamente na borda da delimitação da região de erro do ponto 2, que é equivalente aos picos das gaussianas estarem separados por apenas 11 *pixels*. Como as duas regiões de erro se sobrepõem com valor maior igual a 50% de sobreposição, considera-se como sendo uma redundância de fixações, lembrando que elas são provenientes de pessoas diferentes, pois o equipamento já foi configurado para eliminar derivas de uma mesma fixação a um raio de 10 *pixels*.

Para tratar esses casos de redundância, foram feitas algumas modificações na contabilização das correspondências da métrica de repetibilidade. Para contabilizar uma correspondência, são feitos os seguintes passos:

1. um ponto da imagem I_1 é projetado na imagem I_i, que é uma instância da imagem original.

2. Após a projeção é verificado se o ponto está nos limites da nova imagem, pois em casos de mudança de escala ele pode não estar. Se estiver, segue para o **passo 3**, se não estiver, o ponto é contabilizado como **descartado** e segue de volta para o **passo 1**.

3. É selecionado o ponto da imagem I_i que possui a menor distância para o ponto que foi projetado.

4. É verificado se esse ponto selecionado apresenta uma distância menor que 11 *pixels* para o ponto projetado, o que significa que existe uma probabilidade alta deles serem um par de correspondência, pois a posição do ponto está dentro da faixa de erro de medição do ponto projetado, com uma sobreposição maior que 50%. Se isso ocorrer, é contabilizada uma correspondência. Caso contrário não é contabilizada uma correspondência e a checagem volta para o **passo 1** utilizando um novo ponto da imagem I_1. Se ocorrer uma correspondência a checagem segue para o **passo 5**.

5. Ocorrendo uma correspondência, o ponto da imagem I_i que foi considerado par do ponto projetado é marcado como visitado e não pode ser par de um novo ponto projetado.

6. Todos os pontos da imagem I_1 que são redundantes com relação ao ponto que foi projetado na imagem I_i e que encontrou um correspondente são marcados como **redundantes**, e não serão mais projetados na imagem I_i nas próximas

iterações do algoritmo. Para um ponto ser considerado redundante, ele precisa estar a uma distância menor que 11 *pixels* de outro ponto, ou seja, apresentar mais que 50% de àrea de sobreposição.

7. Na sequência, seleciona-se o próximo ponto da lista de pontos da imagem I_1 que ainda não foi checado e verifica se ele está marcado como **redundante**. Se estiver, seleciona um novo ponto até encontrar um que não seja **redundante** e volta para o **passo 1**. Caso todos os pontos já sejam redundantes ou já tenham sido checados, a busca é finalizada.

Executando esses passos na contabilização de correspondências, é possível eliminar casos críticos de redundância e ter uma representação mais justa de repetibilidade de pontos. Além disso, também foi considerado uma contabilização diferente do total de pontos. Na métrica original, o total de pontos é considerado o menor número de pontos existente entre as duas imagens. No entanto, isso não penaliza completamente as fragilidades do detector, pois o detector ideal é aquele que encontra N pontos na imagem original e N pontos em qualquer instância dessa imagem, claro, considerando que partes da cena não serão eliminadas, mas caso sejam, os pontos que forem projetados fora da imagem deveriam ser considerados descartados e não entrar na contabilização do total de pontos, pois também existem transformações que não eliminam elementos da cena, como borramento e iluminação, por exemplo.

A métrica original sempre considera a menor quantidade de pontos independente de qualquer caso, e isso não aponta completamente as fragilidades do detector, pois se forem encontrados 50 pontos na imagem original e apenas 20 em uma imagem borrada, e esses 20 se repetirem nos lugares corretos, é considerado que o detector é praticamente perfeito e tem 100% de repetibilidade, mas ele não é perfeito, ele é inferior a um detector que nas mesmas condições encontra 50 pontos na imagem original e 40 na imagem borrada, e os 40 pontos se repetem nos locais corretos. Nesse caso, o segundo detector é bem superior ao primeiro, mas a métrica não demonstra isso, pois ela trata os dois como equivalentes com 100% de repetibilidade, o que não é justo.

Para evitar essa contabilização injusta, nesse trabalho, o total de pontos considerado foi dado da seguinte forma:

$$Total = T_{orig} - T_{des} - T_{red} \tag{5.4}$$

Em que T_{orig} é o total de pontos da imagem original, T_{des} é o total de pontos descartados, ou seja, aqueles que quando projetados estavam foram dos limites da

imagem I_i, a depender da transformação, e T_{red} é o total de pontos redundantes.

De acordo com a equação 5.4, o total de pontos é o equivalente a somar os pontos da imagem original que não tiveram correspondência com os pontos que tiveram correspondência. Dessa forma, a métrica de repetibilidade para os mapas de fixação utilizada nesse trabalho é dada por:

$$r_{fix}(\epsilon) = \frac{C}{Total} \tag{5.5}$$

Em que C é o total de correspondências que foram contabilizadas utilizando os passos que foram descritos nesse trabalho.

Definida a métrica de repetibilidade utilizada neste trabalho, a Figura 5.23 apresenta os gráficos de repetibilidade dos pontos, considerando a métrica r_{fix}. A Figura 5.24 ilustra a quantidade de pontos da imagem original que foram considerados correspondências bem sucedidas durante o teste de repetibilidade.

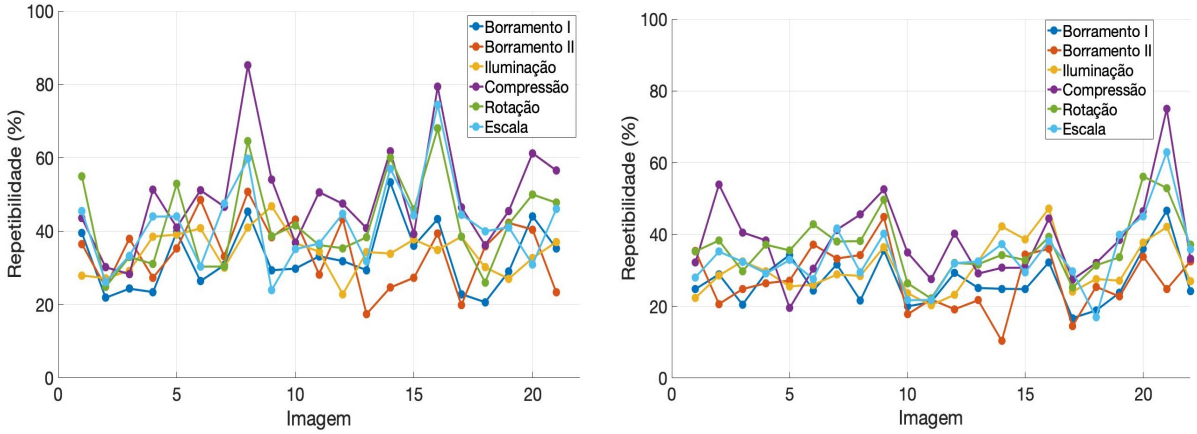


Figura 5.23: Análise de repetibilidade dos pontos de fixação considerando as informações de posição (x,y) dos pontos no mapa. (a) Dados de repetibilidade para imagens do **Grupo I**. (b) Dados de repetibilidade para imagens do **Grupo II**.

Como ilustrado na Figura 5.23, a repetibilidade varia entre 20%, no mínimo, e 90%, no máximo, se mantendo normalmente em torno de 40% na maioria dos casos. Além disso, de acordo com o resultado da métrica, a transformação de borramento I é a que por vezes apresenta um desempenho inferior ao resultado das outras transformações. Já para as transformações de compressão, rotação e escala, os resultados são bem semelhantes. Além disso, de modo geral é observada uma taxa de repetibilidade bastante significativa, pois em várias situações 50% dos pontos da imagem original obtiveram correspondentes nas imagens transformadas, isso sem considerar redundâncias, como foi explicado na definição da métrica, o que é um resultado muito positivo. Também não são observadas diferenças significativas

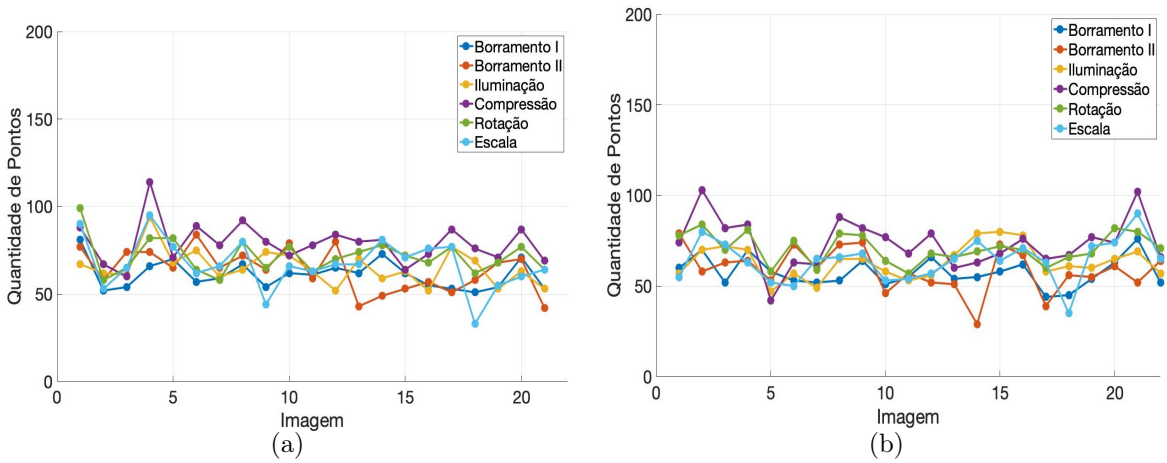


Figura 5.24: Quantidade de pontos contabilizada como correspondências bem sucedidas no teste de repetibilidade considerando as informações de posição (x,y) dos pontos no mapa. (a) Quantidade de pontos considerados correspondentes para imagens do **Grupo I**. (b) Quantidade de pontos considerados correspondentes para imagens do **Grupo II**.

de repetibilidade entre as imagens do grupo I e as imagens do grupo II, o que sugere que o comportamento de invariância do sistema atencional não tem uma dependência direta com o tipo de imagem. Outro fator importante que deve ser considerado é que, os pontos coincidentes são resultantes de voluntários diferentes, e que a nível de fixação há correspondência e repetibilidade dos pontos, ou seja, há consistência entre fixações de indivíduos diferentes em todos os casos que foram testados. Isso significa que pessoas distintas focaram sua retina em micro-regiões da imagem extretamente próximas.

É importante observar que nos casos em que a repetibilidade foi menor, ainda assim, foram obtidos muitos pontos correspondentes. De acordo com a Figura 5.24, que ilustra a quantidade de pontos que foram considerados correspondências bem sucedidas pela métrica, é demonstrado que nos piores casos foram recuperados cerca de 40 pontos, o que é uma quantidade bastante significativa de pontos para qualquer detector local.

Outra observação que deve ser feita é que, os pontos considerados boas correspondências pela métrica não estão conectados, eles estão por vezes próximos, pois fazem parte do mesmo elemento atencional, mas não são redundantes. Isto demonstra que o efeito desejado nas modificações da métrica foi obtido, evitando pontuações redundantes de pontos correspondentes. A Figura 5.25 ilustra, para uma imagem do grupo I, exemplos dos pares de pontos não redundantes que foram selecionados pela métrica e considerados como correspondências bem sucedidas, e a Figura 5.26 ilustra esse resultado para uma imagem do grupo II.

É interessante observar também que, nas duas Figuras, tanto na 5.25 quanto na 5.26, os pontos que possuem pares correspondentes são praticamente os mesmos,

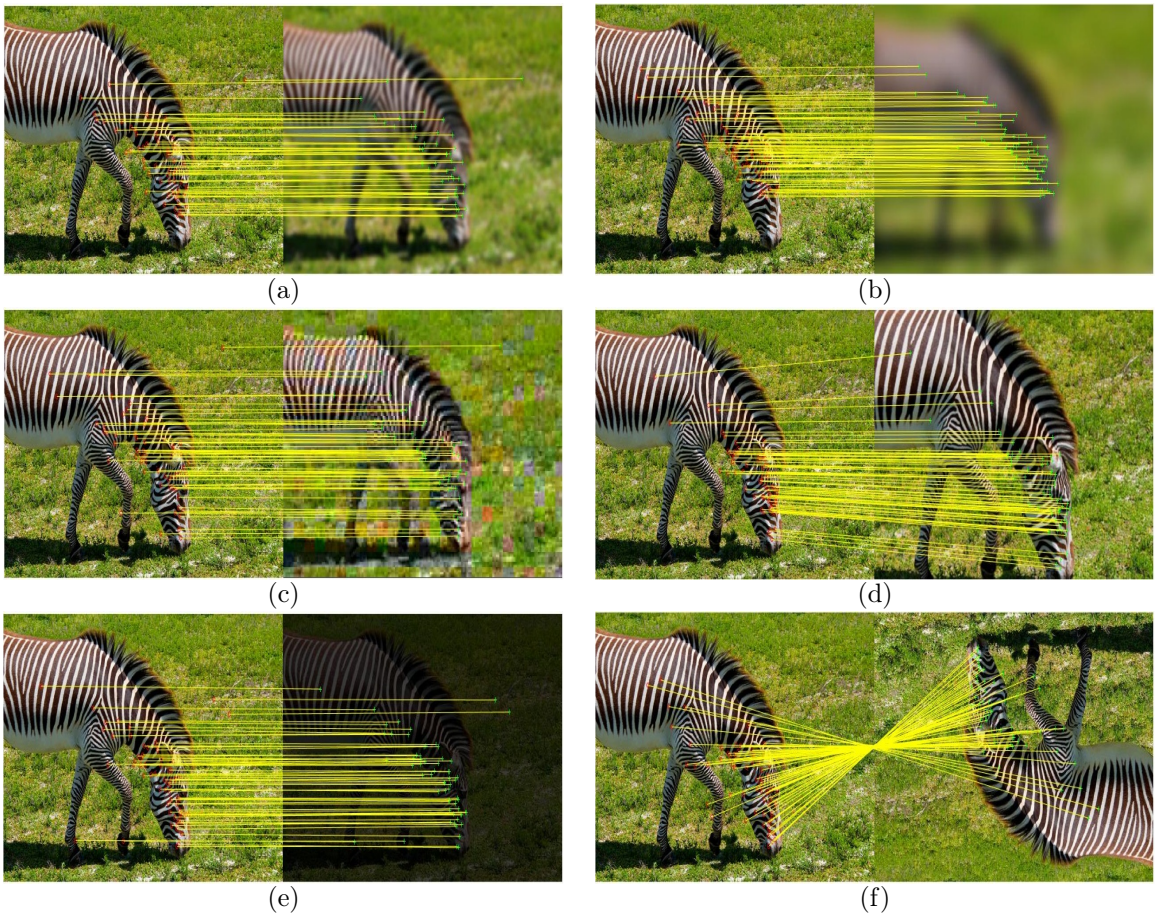


Figura 5.25: Correspondências bem sucedidas consideradas pelo teste de repetibilidade em uma imagem do grupo I. (a) Pares de pontos correspondentes entre a imagem original e a imagem com borramento I. (b) Pares de pontos correspondentes entre a imagem original e a imagem com borramento II. (c) Pares de pontos correspondentes entre a imagem original e a imagem comprimida. (d) Pares de pontos correspondentes entre a imagem original e a imagem escalonada. (e) Pares de pontos correspondentes entre a imagem original e a escurecida. (f) Pares de pontos correspondentes entre a imagem original e a imagem rotacionada.

independente das transformações, ou seja, os pontos da regiões que se repetem na transformação de borramento, de forma geral, se repetem nas outras transformações também, nos mesmos locais, esse resultado é bastante interessante pois é uma demonstração de que a consistência das fixações é bastante semelhante entre as transformações, pois como ilustrado na Figura 5.25 os detalhes dos olhos, do rosto e pescoço da zebra são fixacões persistentes entre as transformações.

Analisada a repetibilidade dos pontos com relação ao seu posicionamento (x,y), é analisado também se o eixo z, o eixo referente ao tempo da fixação, tem influência sobre a repetibilidade, ou seja, se há um padrão de repetibilidade dos pontos também considerando o tempo de fixação. Para isso, foi utilizada a mesma métrica de repetibilidade definida neste trabalho, chamada de r_{fix}, só que ao invés de testar semelhança entre os pontos somente com relação ao eixo (x,y) é testada a semelhança

dos pontos considerando também o tempo de fixação (eixo z). Então, após encontrar o ponto mais próximo no eixo (x,y) e verificar se ele está dentro do raio considerado, é verificado, em seguida, se a diferença entre os valores dos tempos de fixação (eixo z) dos pontos testados é menor ou igual ao valor de um determinado limiar. Foi preciso definir um limiar para realizar a comparação, como não há trabalhos nessa linha, esse limiar, nesse trabalho, foi definido para corresponder a aproximadamente 10% do valor máximo que uma fixação pode assumir, dessa forma o limiar de comparação foi de 0,09. Os resultados obtidos são apresentados nos gráficos das Figuras 5.27 e 5.28.

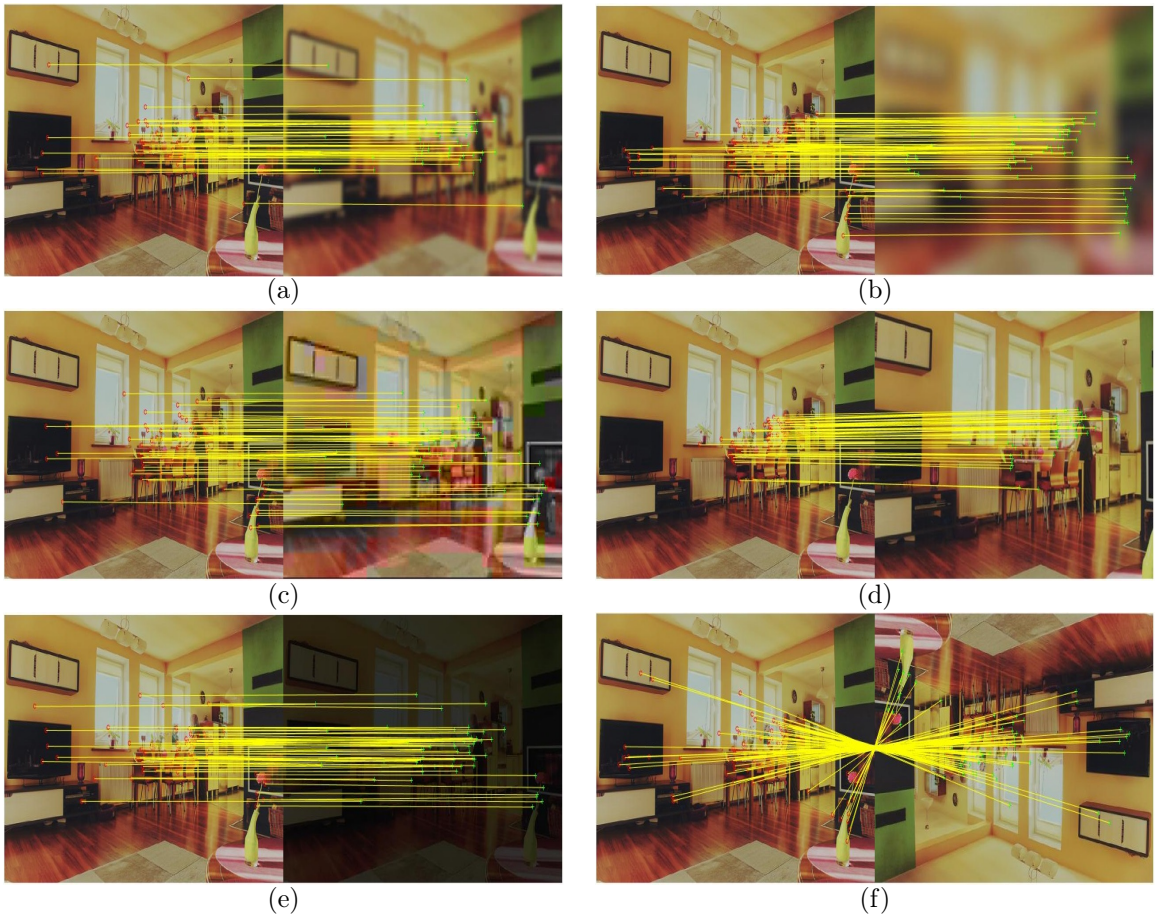


Figura 5.26: Correspondências bem sucedidas consideradas pelo teste de repetibilidade em uma imagem do grupo II. (a) Pares de pontos correspondentes entre a imagem original e a imagem com borramento I. (b) Pares de pontos correspondentes entre a imagem original e a imagem com borramento II. (c) Pares de pontos correspondentes entre a imagem original e a imagem comprimida. (d) Pares de pontos correspondentes entre a imagem original e a imagem escalonada. (e) Pares de pontos correspondentes entre a imagem original e a escurecida. (f) Pares de pontos correspondentes entre a imagem original e a imagem rotacionada.

Esses gráficos mostram diferenças de repetibilidade e de quantidade de correspondências muito pequenas, quando comparados aos resultados obtidos utilizando apenas o eixo (x,y). Isso demonstra um resultado muito importante, pois a repeti-

bilidade não ocorre somente com relação ao posicionamento do ponto, mas também com relação ao tempo em que a retina se fixou nas micro regiões da imagem. O que também demonstra que a informação de saliência, que é representada por esse eixo, é uma informação relevante, que deve ser analisada em detalhes para o desenvolvimento de extratores de características locais melhores.

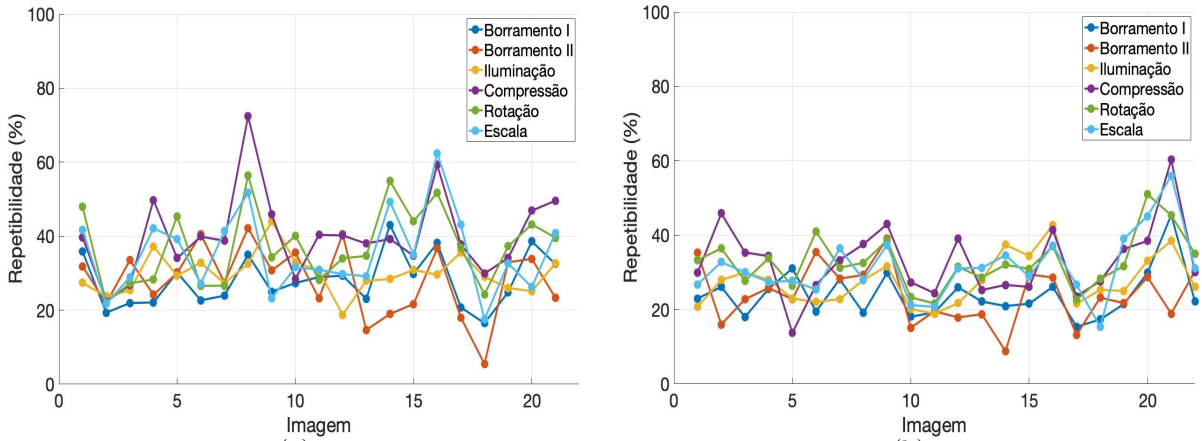


Figura 5.27: Análise de repetibilidade dos pontos de fixação considerando as informações de posição (x,y) e a informação de tempo de fixação cada ponto do mapa. (a) Dados de repetibilidade para imagens do **Grupo I**. (b) Dados de repetibilidade para imagens do **Grupo II**.

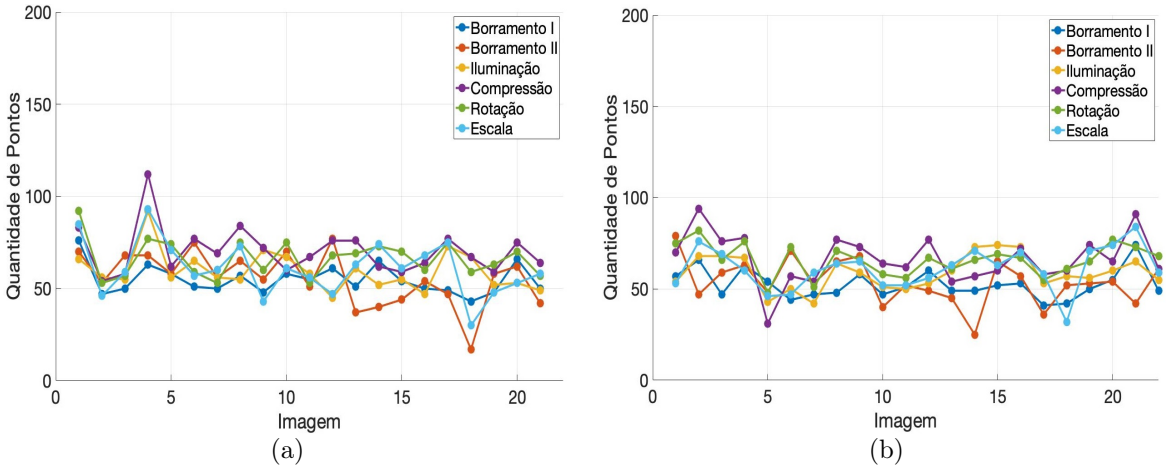


Figura 5.28: Quantidade de pontos contabilizada considerando as informações de posição (x,y) e a informação de tempo de fixação de cada ponto do mapa. (a) Quantidade de pontos considerados correspondentes para imagens do **Grupo I**. (b) Quantidade de pontos considerados correspondentes para imagens do **Grupo II**.

Como discutido nessa seção, os pontos de fixação são bons candidatos a pontos-chave, pois existe repetibilidade considerável nos pontos, existe distinção entre os elementos do mapa quando esse mapa é construído considerando apenas o raio de erro de calibração e quando o mapa e os experimentos são feitos com os devidos

cuidados para evitar o acúmulo de erros. Sob esses critérios, é totalmente viável olhar para mapas de densidade como extratores locais de características. No entanto, os modelos computacionais e os próprios mapas de densidade construídos na área não refletem bem essas características, por isso, a seção 5.2.4 faz uma breve reflexão a respeito da situação atual dos modelos que são estado da arte da área e o que precisa ser modificado nos modelos para que a tarefa de extração de características locais seja possível.

5.2.4 Regiões de Fixação x Modelos Computacionais de Atenção

Já é bem conhecido na área de atenção visual que os modelos que mais se aproximam dos mapas de densidade são os modelos que utilizam *Deep Learning*. Resultados quantitativos mostram isso facilmente no repositório do MIT Benchmark [6]. De certo modo, esses modelos possuem a vantagem de utilizar bases de densidade diretamente como referência nas etapas de treinamento. Além disso, esses modelos contam com redes neurais grandes e com alto poder de extração de características.

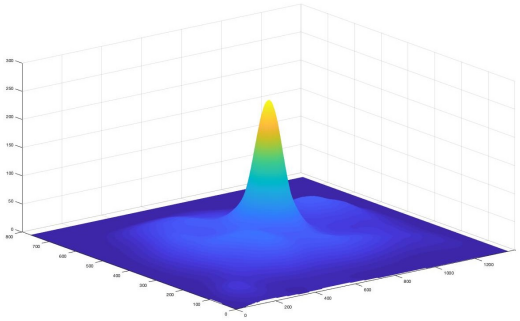
Em média, os maiores valores de similaridade dos mapas de saliência com os mapas de densidade estão em torno de 69%. Isso parece ser um valor elevado, mas na prática ainda há muitas diferenças. A Figura 5.29 em (b) e (c) ilustra a superfície de dois mapas de saliência que são bastante atuais e considerados os melhores modelos da área. Em (d) é apresentado o mapa de densidade construído utilizando o desvio padrão igual a 40, que é o valor de referência da área, e em (e) é apresentado o mapa de densidade construído com o desvio padrão igual a 11, valor utilizado nesse trabalho para construir um mapa de densidade que represente apenas o erro de calibração dos pontos.

Ao comparar as Figuras (b), (c) e (e) já é possível notar muitas diferenças entre as superfícies. A superfície dos dois mapas é muito mais suave e existe menos distinção dos elementos. É como se as gaussianas tivessem sido criadas com um desvio padrão maior que o necessário para representar a densidade dos pontos. Isso faz parecer que o único papel do mapa de saliência é ser um redutor de informação, e que a tarefa de utilizá-lo como extrator de características locais é impossível, já que parece que existe uma gaussiana gigante compondo os mapas de saliência. No entanto, o fato da informação de saliência ser mal representada computacionalmente ainda não significa que é impossível utilizar um modelo de atenção visual para tal tarefa.

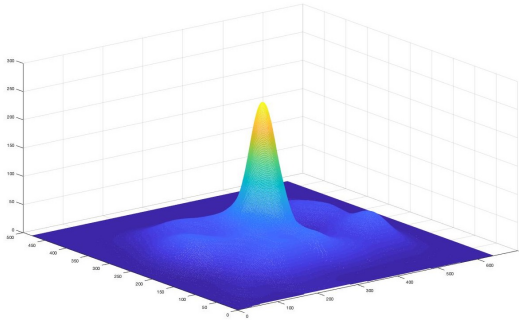
Ao comparar o mapa de densidade clássico com o mapa de densidade modificado é possível notar diferenças. O mapa modificado representa fixações, enquanto o mapa clássico representa a região em ângulo visual correspondente às áreas de maior pico

[6] Pode ser acessado em: `http://saliency.mit.edu/results_mit300.html`

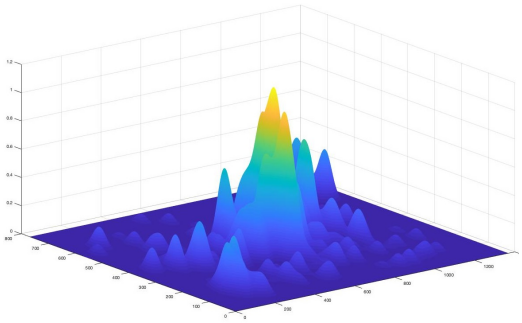
(a)



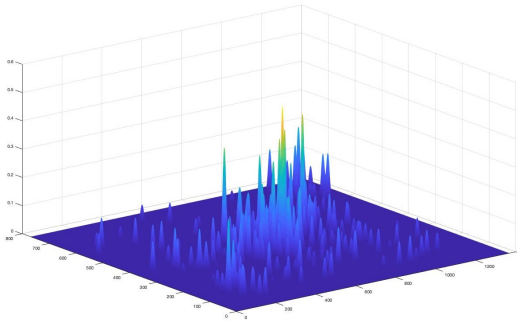
(b)



(c)



(d)



(e)

Figura 5.29: Mapa de atenção com valores de σ diferentes. (a) Imagem original. (b) Superfície de saliência do modelo Sam-ResNet. (c) Superfície de saliência do modelo Sam-VGG. (d) Superfície tradicional construída a partir das fixacões, com $\sigma = 40$. (e) Superfície de densidade das fixações, com $\sigma = 11$.

de absorção de informação por parte da fóvea. Novamente, essa representação é bem coerente, mas causa vários empecilhos e impossibilita o trabalho proposto, pois essa representação já camufla várias fixações.

Os modelos que não utilizam aprendizado de máquina são ainda bem menos comportados e estão ainda mais distantes de alcançar potencial como extratores locais de características. Os modelos que podem chegar mais perto dessa habilidade, por enquanto, são os que utilizam técnicas mais avançadas de aprendizado de máquina, tais como o Sam-VGG e o Sam-ResNet. Isso é afirmado pois foi observado que, durante a fase de treino, alguns desses modelos utilizam também os dados das fixações e após se aproximar da localização delas atribuem as gaussianas, só que essas gaussianas ainda não são bem definidas. O Sam-Resnet, por exemplo, utiliza 16 gaussianas, mesmo encontrando mais fixações. Além disso, o desvio padrão dessas gaussianas é controlado pela fase de treino. No entanto, para começar a modificar esses modelos é preciso que sejam construídas bases de mapas de densidade ainda maiores que a base que foi construída nesse trabalho.

Então, na verdade, o que impossibilita o uso dos modelos computacionais de atenção para extração de características locais são dois fatores principais: ausência de uma investigação a respeito do potencial de invariância do sistema atencional humano e ausência de mapas de densidade que representem essas fixações, para que os modelos computacionais possam segui-lo.

Capítulo 6

Conclusões

Nesse trabalho foi apresentada uma investigação a respeito de sistemas atencionais de visão como extratores de características locais. Inicialmente, foi realizada uma investigação utilizando os modelos computacionais mais próximos da resposta dos mapas de densidade, que são os mapas de referência para a área de atenção visual. Foram escolhidos cinco modelos, dois modelos que são tidos como estado da arte da área, dois modelos que apresentam também bons resultados, mas são mais antigos, e por fim, o modelo clássico proposto por Itti e Koch.

A investigação nos modelos computacionais revelou vários problemas considerados críticos para admitir modelos de atenção como extratores de características locais. Os principais problemas encontrados foram a falta de distinção entre os pontos, associada à baixa dimensionalidade em que os mapas são construídos. Além disso, a falta de padrão na resposta das superfícies de saliência, quando testadas com imagens que sofreram transformações, é muito alta. As superfícies de atenção computacionais são muito suaves, então o gradiente não é importante nessas superfícies e ele não diz muito a respeito das regiões de saliência, já que a maioria das áreas apresenta grandes *platôs*. Ainda com todas essas observações, foram tentadas técnicas de seleção de máximos globais, técnicas de agrupamento de pontos e técnicas de filtragem de pontos utilizando filtros clássicos do processamento de imagens, mas em nenhuma das tentantivas foram obtidos resultados satisfatórios.

Esses resultados insatisfatórios foram atribuidos à existência de um problema principal na investigação: **a falta de um critério de invariância para os pontos de saliência**. Essa falta de critério torna a busca cega e extremamente difícil, mas, a partir de tudo que foi testado e pelo fato dos mapas de saliência não serem claramente construídos com características de invariância, infere-se que essa tarefa nos modelos computacionais ainda não é factível.

Por isso, a investigação desse trabalho seguiu para os mapas de densidade, pois eles são a melhor fonte de dados para investigar se é possível extrair pontos invariantes de sistemas de atenção visual. No entanto, não há bases de mapas de fixação

construídas para estudar esse aspecto. Para resolver essa dificuldade e continuar a investigação do trabalho, foi construída uma base de mapas de fixação com 313 voluntários distintos. Essa base possui 6 instâncias da imagem original, sendo que cada instância representa uma transformação diferente.

Com a construção da base foi possível notar que os mapas de densidade trazem uma representação camuflada dos pontos de fixação. Essa camuflagem é feita por gaussianas e tem como principal objetivo representar a área de maior absorção de informação da fóvea humana. Para esse trabalho, foi criada uma representação diferente, atribuindo gaussianas mais finas aos pontos de fixação, representando o ponto apenas com o erro de calibração. Esse erro foi devidamente calculado de acordo com as condições em que o experimento foi realizado. Todos esses detalhes resultaram em uma representação bem diferente dos pontos de fixação. Nessa representação existe distinção entre os pontos e repetibilidade.

Durante os experimentos foi notado que os pontos de fixação seriam os potenciais candidatos a pontos-chave dentro de um mapa de saliência. O comportamento desses pontos foi avaliado qualitativamente e quantitativamente. Qualitativamente foi observado que a consistência das fixações se mantém, apesar das transformações locais ocorridas em uma imagem. Claro que existem modificações sutis entre elas. Além disso, foram observados padrões de observação que os seres humanos possuem que podem estar associados a instintos biológicos e comportamentos culturais.

Para avaliar quantitavamente o potencial de invariância dos pontos de fixação, foram realizadas modificações na métrica clássica de repetibilidade, com o objetivo de construir um resultado quantitativo mais adequado às características dos pontos de fixação. Com relação a repetibilidade, foram obtidos resultados entre 20% e 90%, e foram encontrados muitos pontos com correspondências bem sucedidas. Além disso, os resultados de repetibilidade não foram fortemente alterados quando a comparação foi realizada utilizando o tempo das fixações dos mapas de densidade, o que demonstra que existe uma consistência das fixações não só com relação a posição (x,y) dos pontos, mas que o tempo das fixações também desempenha um papel importante.

Todos esses fatores corroboram o fato de que existe invariância no sistema atencional humano, até mesmo quando ele é submetido a analisar imagens que estão sob condições que não são naturais, como imagens borradas e rotacionadas a 180°, por exemplo. Os resultados mostraram um ótimo comportamento dos pontos de fixação nas imagens rotacionadas, pois era esperado que nessas imagens as fixações fossem encontradas em locais diferentes, já que as imagens estão invertidas. No entanto, as fixações foram encontradas nos mesmos locais que as fixações das imagens originais.

Também foi realizada uma pequena avaliação comparativa dos dados das fixações com alguns dos principais detectores clássicos, e foram observados resultados dis-

tintos, o que demonstra que a atenção humana não é somente guiada por gradiente local, mas que existem outros elementos na imagem, e a composição de elementos guia a atenção. Além disso, é demonstrado o potencial dos mapas de fixação, não só como elementos de referência para sistemas atencionais, mas para os próprios detectores clássicos.

Com a construção da base e com os resultados obtidos, foi possível concluir também que os modelos computacionais são construídos ainda de uma forma muito distante dos mapas de densidade tradicionais, e que 69% de similaridade com os mapas de densidade ainda é muito pouco quando o objetivo é analisar os mapas sob o ângulo de extratores de características. Além disso, foi observado que os mapas de densidade também possuem uma representação que não contribui para a ánalise desse trabalho, pois os pontos de fixação perdem parte da sua distinção e se torna mais difícil analisar os mapas. Esse aspecto também justifica a importância da construção de uma base própria para os objetivos desse trabalho, pois sem a construção passo a passo dos mapas de densidade, não seria possível identificar esses fatores de representação de modelo.

Os resultados trazidos por esse trabalho apontam novos horizontes para trabalhos futuros em várias direções. Sendo esses os principais deles:

- **Construção de novas bases de mapas de densidade:** O trabalho atual teve como objetivo construir uma base simples e pequena com apenas 6 instâncias da imagem original. Foram observados bons resultados, mas transformações mais complexas não foram testadas. Seria interessante construir bases ainda maiores para extrair mais conclusões a respeito do assunto, e também para fornecer mapas de densidade invariantes que possam ser utilizados como referência para construção de mapas de saliência.

- **Criação de Modelos representativos:** Foi feita uma pequena modificação na representação dos pontos de fixação e já foi possível extrair muitas conclusões que não eram possíveis anteriormente. Mas a representação atual possui pontos redundantes, o que foi tratado pelas modificações da métrica de repetibilidade, para evitar resultados que não reflitam bem o comportamento real dos pontos. Seria interessante construir uma representação mais limpa, que elimine as fixações redundantes mas que mantenha o nível de distinção do mapa de densidade. Uma representação mais limpa e clara de quais são os pontos mais consistentes é mais interessante para os modelos computacionais utilizarem como base.

- **Modificações e Aperfeiçoamentos nos detectores clássicos:** Como foi discutido na Seção 5.2.1, a resposta dos detectores clássicos ainda é bem diferente das fixações, ou seja, ainda existem aspectos a serem investigados e os

próprios detectores clássicos podem ser aprimorados, utilizando como base os dados das fixações humanas.

- **Construção e Adaptação dos Modelos Computacionais de Atenção:** Com bases de mapas de densidade apropriadas para a tarefa de extração de caracterísicas, os modelos computacionais de atenção podem ser mais que simples delimitadores de área. Porém, é necessário realizar modificações e atualizações nos modelos computacionais para que eles possam funcionar como extratores de características.

- **Avaliação dos descritores clássicos:** Os descritores clássicos são em sua maior parte construídos para lidar com gradiente local. Se os mapas de fixação podem ser considerados uma representação de um detector, então é viável avaliar o desempenho dos descritores clássicos quando utilizados para descrever pontos de fixação. Podem ser analisadas as vantagens e desvantagens de cada abordagem e podem surgir novas modificações nos descritores clássicos, ou até mesmo novos descritores.

Todos os itens citados se referem a uma porção de caminhos de investigação que podem ser seguidos a partir dos resultados obtidos nesse trabalho. Há muitas áreas interessadas nesses resultados, como por exemplo, a área de recuperação de imagem por conteúdo. Nessa área, são necessários detectores locais que sejam compactos e que representem bem o conteúdo da cena. Detectores baseados nos mapas de fixação podem atender bem essas necessidades. Além disso, novos detectores e descritores utilizando os dados da base desenvolvida nesse trabalho podem ser promissores em inúmeras aplicações da visão computacional.

Referências Bibliográficas

[1] BENICASA, A. X. *Sistemas computacionais para atenção visual Top-Down e Bottom-up usando redes neurais artificiais*. Tese de Doutorado, Universidade de São Paulo.

[2] SIAGIAN, C., ITTI, L. "Biologically inspired mobile robot vision localization", *Robotics, IEEE Transactions on*, v. 25, n. 4, pp. 861–873, 2009.

[3] ITTI, L., KOCH, C., NIEBUR, E. "A model of saliency-based visual attention for rapid scene analysis", *IEEE Transactions on pattern analysis and machine intelligence*, v. 20, n. 11, pp. 1254–1259, 1998.

[4] FANG, S., LI, J., TIAN, Y., et al. "Learning discriminative subspaces on random contrasts for image saliency analysis", *IEEE transactions on neural networks and learning systems*, v. 28, n. 5, pp. 1095–1108, 2017.

[5] CORNIA, M., BARALDI, L., SERRA, G., et al. "Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model." *CoRR*, v. abs/1611.09571, 2016. Disponível em: <http://dblp.uni-trier.de/db/journals/corr/corr1611.html#CorniaBSC16a>.

[6] LUO, N., SUN, Q., CHEN, Q., et al. "A Novel Tracking Algorithm via Feature Points Matching", *PloS one*, v. 10, n. 1, pp. e0116315, 2015.

[7] JUDD, T., DURAND, F., TORRALBA, A. "Fixations on low resolution images", *Journal of Vision*, v. 10, n. 7, pp. 142–142, 2010.

[8] SCHMID, C., MOHR, R., BAUCKHAGE, C. "Evaluation of interest point detectors", *International Journal of computer vision*, v. 37, n. 2, pp. 151–172, 2000.

[9] MACHADO, A. M. C. *Metodologia para Reconhecimento de Padrões em Visão Computacional*. Tese de Doutorado, dissertação de mestrado submetida a Universidade Federal de Minas Gerais, 1994.

[10] MARR, D. "Vision: A computational approach". 1982.

[11] ITTI, L., KOCH, C. "Computational modelling of visual attention", *Nature reviews neuroscience*, v. 2, n. 3, pp. 194–203, 2001.

[12] ITTI, L., KOCH, C. "A saliency-based search mechanism for overt and covert shifts of visual attention", *Vision research*, v. 40, n. 10-12, pp. 1489–1506, 2000.

[13] BORJI, A., ITTI, L. "State-of-the-art in visual attention modeling", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 35, n. 1, pp. 185–207, 2013.

[14] JUDD, T., DURAND, F., TORRALBA, A. "A benchmark of computational models of saliency to predict human fixations", 2012.

[15] ITTI, L. *Models of bottom-up and top-down visual attention*. Tese de Doutorado, California Institute of Technology, 2000.

[16] LOWE, D. G. "Distinctive image features from scale-invariant keypoints", *International journal of computer vision*, v. 60, n. 2, pp. 91–110, 2004.

[17] BAY, H., ESS, A., TUYTELAARS, T., et al. "Speeded-up robust features (SURF)", *Computer vision and image understanding*, v. 110, n. 3, pp. 346–359, 2008.

[18] TSOTSOS, J. K., CULHANE, S. M., WAI, W. Y. K., et al. "Modeling visual attention via selective tuning", *Artificial Intelligence*, v. 78, pp. 507–545, 1995.

[19] MILLER, E. K. "THE PREFRONTAL CORTEX AND COGNITIVE CONTROL", *NATURE REVIEWS| NEUROSCIENCE*, v. 1, pp. 59, 2000.

[20] RODRIGUES, F. A. *Localização e reconhecimento de placas de sinalização utilizando um mecanismo de atenção visual e redes neurais artificiais*. Tese de Doutorado, Universidade Federal de Campina Grande, 2002.

[21] CORBETTA, M. "Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems?" *Proceedings of the National Academy of Sciences*, v. 95, n. 3, pp. 831–838, 1998.

[22] FRINTROP, S., ROME, E., CHRISTENSEN, H. I. "Computational visual attention systems and their cognitive foundations: A survey", *ACM Transactions on Applied Perception (TAP)*, v. 7, n. 1, pp. 6, 2010.

[23] TREISMAN, A. M., GELADE, G. "A feature-integration theory of attention", *Cognitive psychology*, v. 12, n. 1, pp. 97–136, 1980.

[24] HAREL, J., KOCH, C., PERONA, P. "Graph-based visual saliency". In: *Advances in neural information processing systems*, pp. 545–552, 2007.

[25] KRUIZINGA, P., PETKOV, N. "Nonlinear operator for oriented texture", *Image Processing, IEEE Transactions on*, v. 8, n. 10, pp. 1395–1407, 1999.

[26] ITTI, L., KOCH, C. "Comparison of feature combination strategies for saliency-based visual attention systems". In: *Electronic Imaging'99*, pp. 473–482. International Society for Optics and Photonics, 1999.

[27] ITTI, L., KOCH, C. "Feature combination strategies for saliency-based visual attention systems", *Journal of Electronic Imaging*, v. 10, n. 1, pp. 161–169, 2001.

[28] ITTI, L., KOCH, C., NIEBUR, E. "A model of saliency-based visual attention for rapid scene analysis", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , n. 11, pp. 1254–1259, 1998.

[29] DE ARAUJO, W. O., COELHO, C. J. *Análise de componentes principais (PCA)*. Relatório técnico, Technical report, UniEvangélica, 2009.

[30] RODRIGUES, C. F. "Análise comparativa entre os métodos decomposição em valores singulares e análise de componentes principais envolvendo matrizes esparsas de grande porte", 2011.

[31] GOODFELLOW, I., BENGIO, Y., COURVILLE, A., et al. *Deep learning*, v. 1. MIT press Cambridge, 2016.

[32] BYLINSKII, Z., JUDD, T., OLIVA, A., et al. "What do different evaluation metrics tell us about saliency models?" *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[33] JUDD, T., DURAND, F., TORRALBA, A. "A benchmark of computational models of saliency to predict human fixations", 2012.

[34] TRUCCO, E., VERRI, A. *Introductory techniques for 3-D computer vision*, v. 201. Prentice Hall Englewood Cliffs, 1998.

[35] LOWE, D. G. "Object recognition from local scale-invariant features". In: *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, v. 2, pp. 1150–1157. Ieee, 1999.

[36] TUYTELAARS, T., MIKOLAJCZYK, K., OTHERS. "Local invariant feature detectors: a survey", *Foundations and trends® in computer graphics and vision*, v. 3, n. 3, pp. 177–280, 2008.

[37] ATTNEAVE, F. "Some informational aspects of visual perception." *Psychological review*, v. 61, n. 3, pp. 183, 1954.

[38] SMITH, S. M., BRADY, J. M. "SUSAN—a new approach to low level image processing", *International journal of computer vision*, v. 23, n. 1, pp. 45–78, 1997.

[39] ROSTEN, E., DRUMMOND, T. "Machine learning for high-speed corner detection". In: *European conference on computer vision*, pp. 430–443. Springer, 2006.

[40] MAIR, E., HAGER, G. D., BURSCHKA, D., et al. "Adaptive and generic corner detection based on the accelerated segment test". In: *European conference on Computer vision*, pp. 183–196. Springer, 2010.

[41] BAY, H., TUYTELAARS, T., VAN GOOL, L. "Surf: Speeded up robust features". In: *European conference on computer vision*, pp. 404–417. Springer, 2006.

[42] VIOLA, P., JONES, M. J. "Robust real-time face detection", *International journal of computer vision*, v. 57, n. 2, pp. 137–154, 2004.

[43] DONOSER, M., BISCHOF, H. "Efficient maximally stable extremal region (MSER) tracking". In: *null*, pp. 553–560. IEEE, 2006.

[44] AGRAWAL, M., KONOLIGE, K., BLAS, M. R. "Censure: Center surround extremas for realtime feature detection and matching". In: *European Conference on Computer Vision*, pp. 102–115. Springer, 2008.

[45] YU, G., MOREL, J.-M. "ASIFT: An algorithm for fully affine invariant comparison", *Image Processing On Line*, v. 1, pp. 11–38, 2011.

[46] KADIR, T., BRADY, M. "Saliency, scale and image description", *International Journal of Computer Vision*, v. 45, n. 2, pp. 83–105, 2001.

[47] KADIR, T., BRADY, M. "Scale saliency: A novel approach to salient feature and scale selection", 2003.

[48] LENC, K., VEDALDI, A. "Learning covariant feature detectors". In: *European Conference on Computer Vision*, pp. 100–117. Springer, 2016.

[49] ZHANG, X., FELIX, X. Y., KARAMAN, S., et al. "Learning Discriminative and Transformation Covariant Local Feature Detectors." In: *CVPR*, pp. 4923–4931, 2017.

[50] YI, K. M., TRULLS, E., LEPETIT, V., et al. "Lift: Learned invariant feature transform". In: *European Conference on Computer Vision*, pp. 467–483. Springer, 2016.

[51] MIKOLAJCZYK, K., SCHMID, C. "A performance evaluation of local descriptors", *IEEE transactions on pattern analysis and machine intelligence*, v. 27, n. 10, pp. 1615–1630, 2005.

[52] KE, Y., SUKTHANKAR, R. "PCA-SIFT: A more distinctive representation for local image descriptors". In: *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, v. 2, pp. II–II. IEEE, 2004.

[53] DALAL, N., TRIGGS, B. "Histograms of oriented gradients for human detection". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, v. 1, pp. 886–893. IEEE, 2005.

[54] BOSCH, A., ZISSERMAN, A., MUÑOZ, X. "Scene classification using a hybrid generative/discriminative approach", *IEEE transactions on pattern analysis and machine intelligence*, v. 30, n. 4, pp. 712–727, 2008.

[55] VAN DE WEIJER, J., GEVERS, T., BAGDANOV, A. D. "Boosting color saliency in image feature detection", *IEEE transactions on pattern analysis and machine intelligence*, v. 28, n. 1, pp. 150–156, 2006.

[56] VAN DE SANDE, K., GEVERS, T., SNOEK, C. "Evaluating color descriptors for object and scene recognition", *IEEE transactions on pattern analysis and machine intelligence*, v. 32, n. 9, pp. 1582–1596, 2010.

[57] ZHU, C., BICHOT, C.-E., CHEN, L., et al. "Visual object recognition using DAISY descriptor". In: *2011 IEEE International Conference on Multimedia and Expo (ICME 2011)*, pp. 1–6. IEEE, 2011.

[58] HUANG, D., ZHU, C., BICHOT, C.-E., et al. "HSOG: a novel local descriptor based on histograms of second order gradients for object categorization". In: *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pp. 199–206. ACM, 2013.

[59] TUYTELAARS, T., SCHMID, C. "Vector quantizing feature space with a regular lattice". In: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8. IEEE, 2007.

[60] WINDER, S., HUA, G., BROWN, M. A. "Picking the best daisy", 2009.

[61] CALONDER, M., LEPETIT, V., FUA, P., et al. "Compact signatures for high-speed interest point description and matching". In: *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 357–364. IEEE, 2009.

[62] SHAKHNAROVICH, G. *Learning task-specific similarity*. Tese de Doutorado, Massachusetts Institute of Technology, 2005.

[63] GIONIS, A., INDYK, P., MOTWANI, R., et al. "Similarity search in high dimensions via hashing". In: *Vldb*, v. 99, pp. 518–529, 1999.

[64] CALONDER, M., LEPETIT, V., OZUYSAL, M., et al. "BRIEF: Computing a local binary descriptor very fast", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 34, n. 7, pp. 1281–1298, 2012.

[65] LEUTENEGGER, S., CHLI, M., SIEGWART, R. Y. "BRISK: Binary robust invariant scalable keypoints". In: *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2548–2555. IEEE, 2011.

[66] PIMENOV, V. "Fast image matching with visual attention and SURF descriptors". In: *Proceedings of the 19th International Conference on Computer Graphics and Vision*, pp. 49–56, 2009.

[67] ALAHI, A., ORTIZ, R., VANDERGHEYNST, P. "Freak: Fast retina keypoint". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 510–517. Ieee, 2012.

[68] TRZCINSKI, T., CHRISTOUDIAS, M., FUA, P., et al. "Boosting binary keypoint descriptors". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2874–2881, 2013.

[69] SCHAPIRE, R. E. "Explaining adaboost". In: *Empirical inference*, Springer, pp. 37–52, 2013.

[70] IWAMOTO, K., MASE, R., NOMURA, T. "BRIGHT: A scalable and compact binary descriptor for low-latency and high accuracy object identification". In: *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pp. 2915–2919. IEEE, 2013.

[71] LEVI, G., HASSNER, T. "LATCH: learned arrangements of three patch codes". In: *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–9. IEEE, 2016.

[72] LIN, K., LU, J., CHEN, C.-S., et al. "Learning compact binary descriptors with unsupervised deep neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1183–1192, 2016.

[73] LIU, Z., LI, Z., ZHANG, J., et al. "Euclidean and hamming embedding for image patch description with convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 72–78, 2016.

[74] MADEO, S., BOBER, M. "Fast, compact, and discriminative: Evaluation of binary descriptors for mobile applications", *IEEE Transactions on Multimedia*, v. 19, n. 2, pp. 221–235, 2017.

[75] BYLINSKII, Z., JUDD, T., OLIVA, A., et al. "What do different evaluation metrics tell us about saliency models?" *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[76] CORREIA, A. S., FREIRE, E. O., MOLINA, L. "Aplicação de mapas de saliência como limitadores de região para detectores clássicos na tarefa de odometria visual." *XXII Congresso Brasileiro de Automática*, 2018.

[77] BONN, R. F.-W.-U. "VOCUS: A Visual Attention System for Object Detection and Goal-directed Search", 2006.

[78] GAO, K., LIN, S., ZHANG, Y., et al. "Attention model based sift keypoints filtration for image retrieval". In: *Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference on*, pp. 191–196. IEEE, 2008.

[79] LIANG, Z., FU, H., CHI, Z., et al. "Salient-sift for image retrieval". In: *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 62–71. Springer, 2010.

[80] LÓPEZ-GARCÍA, F., FDEZ-VIDAL, X. R., PARDO, X. M., et al. "Scene recognition through visual attention and image features: A comparison between sift and surf approaches". In: *Object Recognition*, InTech, 2011.

[81] MESQUITA, R. G. D. "Reconhecimento de instâncias guiado por algoritmos de atençao visual", 2017.

[82] HEINEN, M. R., ENGEL, P. M. "NLOOK: a computational attention model for robot vision", *Journal of the Brazilian Computer Society*, v. 15, n. 3, pp. 3–17, 2009.

[83] NEWMAN, P., HO, K. "SLAM-loop closing with visually salient features". In: *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pp. 635–642. IEEE, 2005.

[84] FRINTROP, S. "The high repeatability of salient regions". In: *Workshop on Vision in Action: Efficient strategies for cognitive agents in complex environments*, 2008.

[85] BERNSEN, J. "Dynamic thresholding of grey-level images", *Proc. 8th International Conference on Pattern Recognition*, pp. 1251–1255, 1986.

[86] GONZALEZ, R. C., WOODS, R. E. *Processamento de imagens digitais*. Edgard Blucher, 2000.

[87] OOMS, K., DUPONT, L., LAPON, L., et al. "Accuracy and precision of fixation locations recorded with the low-cost Eye Tribe tracker in different experimental setups", *Journal of eye movement research*, v. 8, n. 1, 2015.

[88] JOHANSSON, J., SOLLI, M., MAKI, A. "An evaluation of local feature detectors and descriptors for infrared images". In: *European Conference on Computer Vision*, pp. 711–723. Springer, 2016.

[89] BYLINSKII, Z., JUDD, T., OLIVA, A., et al. "What do different evaluation metrics tell us about saliency models?" *IEEE transactions on pattern analysis and machine intelligence*, 2018.